DOCUMENT RESUME

ED 428 084 TM 029 484

AUTHOR	Brooks, Gordon P.
TITLE	Perry: Fact, Fiction, and Outcomes Assessment.
PUB DATE	1998-10-00
NOTE	78p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Adolescents; *Cognitive Development; Developmental Stages; Educational Assessment; *Educational Theories; Outcomes of Education; Qualitative Research; *Research Design; Scaling; *Validity
IDENTIFIERS	*Perry Developmental Scheme; Perry (William G Jr)

ABSTRACT

Theories concerning cognitive development or complexity include one developed by W. G. Perry, Jr., and his colleagues (1970). Research related to Perry's scheme is analyzed to suggest ways by which this particular theory can be validated. The first part of the paper describes Perry's scheme and assumptions of cognitive development theories in general. Perry described late-adolescent development in terms of nine positions, or coherent forms of thought. The scheme represents a continuum that describes the steps by which students move from a simplistic, categorical view of the world to a realization of the contingent nature of knowledge and values to the formation and affirmation of their own commitments. Other researchers have attempted to operationalize the relevant constructs in Perry's theory. The evidence concerning the reliability of currently used measures based on Perry's scheme is reviewed, including information on scaling and scoring. Other alternatives for measuring these relevant constructs in Perry's theory are explored. A look at all these approaches suggests some ways in which Perry's theory might be validated. Accountability issues in student personnel are of increasing interest, and the Perry scheme has considerable relevance for these concerns. Some of the major issues involved in developmental research are reviewed, including assessment problems, measurement issues, selection of sample subjects, and research design difficulties. The design of a course of research is explored, including a discussion of the relevance of qualitative methods to such research, improvements on efforts by others, and the possibility of alternative scaling procedures. (Contains 2 tables, 1 figure, and 114 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TM

Running Head: PERRY ASSESSMENT

ED 428 084

Perry: Fact, fiction, and outcomes assessment

Gordon P. Brooks
Ohio University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Gordon Brooks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

TM029484

Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL (1998, October)

BEST COPY AVAILABLE

THE VALIDATION OF THE PERRY SCHEME

Theories concerning cognitive development or complexity include the one of Perry (1970) and his colleagues. Research testing or validating this particular theory is reasonably scarce in the literature. The ultimate purpose of this paper is to recommend a means by which Perry's scheme can be validated. Along the way, several critical areas of the literature will be analyzed, including (a) Perry's theory and its assumptions, (b) some implications of Perry's scheme, and (c) how other researchers have operationalized the scheme. Finally, some recommendations for further research will be presented by which Perry's theory can be examined critically, including such areas as instrumentation, research design, subjects, and limitations of such an undertaking.

The Perry Scheme of Intellectual and Ethical Development

The paper begins with a description of Perry's theory, including assumptions of cognitive development theories in general. Relevant papers will be cited, examples will be provided, and limitations and implications will be indicated.

Assumptions of Cognitive Developmental Theories

Rodgers (1989) argued that "intellectual development is central to the primary purposes of higher education in the English, German, and American traditions.... It is the one form of development in which faculty and student affairs staff may have a common commitment, and it has wide application to practice in both domains" (p. 143). Historically, intellectual and ethical development have been major goals of higher education in the attempt to produce well-rounded graduates (Moore, 1989).

Cognitive developmental theories attempt to describe the development of how people make sense out of their experiences. Basically, cognitive developmental theories outline the successive changes in how people think and the increasing complexity with which they make meaning of questions of knowledge, morality, valuing, faith, or self (Rodgers, 1990; Widick, Knefelkamp, & Parker, 1980). Individuals interpret, or make meaning of, their educational experience as a result of their assumptions about the nature, limits, and certainty of their knowledge. The concepts and assumptions of the theoretical tradition of cognitive development were articulated first by Piaget (King, 1978; Widick, Knefelkamp, & Parker, 1980).

First, cognitive developmental theories assume an information processing view of the individual (King, 1978). That is, individuals actively interpret their environment using cognitive structures in order to perceive, organize, evaluate, and thereby make meaning from their experiences. Cognitive structures refer to the mediating cognitive filters, or interpretive frameworks, that help determine how a person translates and interacts with external reality. Cognitive structures are the assumptions about knowledge, truth, and reality that we use to "shape the way we see the world and ourselves as participants in it" (Belenky, Clinchy, Goldberger, & Tarule, 1986, p. 3). Generally, epistemological growth evolves from a concrete structural assumption that knowledge is absolute to an abstract assumption that knowledge may be uncertain to a relativistic assumption that forms a probabilistic view of knowledge (Kitchener, 1982). For college student development, Moore and Hunter (1993) specified that the student's role changes from passive learner of facts and truths to active creator of argument and knowledge.

Second, cognitive development is interactive. Development occurs as a result of the interaction between the individual and the environment. People need to make sense of their experiences, that is, to interpret them in a meaningful manner (Perry, 1970). How individuals interpret an event or issue will differ based on the cognitive structure they bring to the experience. A situation that is congruent with an individual's previous experiences will require little effort in order for the individual to make sense of it. Through selection, simplification, and distortion, an individual assimilates the new experience into the cognitive forms brought to the situation. When assimilation of new information occurs, the individual remains in relative equilibrium at the current way of thinking.

However, when an idea, problem, or situation bears little resemblance to prior experiences, cognitive conflict occurs within the individual. This confusion requires recombinations, reorganizations, and transformations of the new and old experiences so that new meanings can emerge. The manifestation of this process of accommodation is often sensed as an insight or realization (Perry, 1970). When accommodation becomes pervasive enough in the interactions between the individual and the environment, disequilibrium occurs and the cognitive structure is altered to handle more complexity. Perry (1970) wrote that "one form leads to another through differentiations and reorganizations required for the meaningful interpretation of increasingly complex experience" (p. 3).

Third, development is seen as a progression along a hierarchical sequence of invariant and irreversible stages, with each stage representing a qualitatively different way of thinking or of approaching problems (King, 1978; Knefelkamp & Slepitza, 1976). Each stage is seen as a more complex set of stable and coherent cognitive structures that subsumes all previous stages (Brabeck, 1984). The movement to each successive position is considered "development rather than mere changes" (Perry, 1981, p. 78) because each position both includes and transcends earlier ones; development therefore proceeds through stages and transitions toward cognitive structures of greater complexity.

Although development is sequential, no assumption is made regarding the intervals between the stages or time spent within a stage; that is, development may proceed at an irregular rate and may occur in spurts (Perry, 1981; Widick, 1977). Because the stages are typically described as invariant and therefore not bound by culture, everyone is believed to pass through each stage in the same order. After attaining a stage, however, an individual may have limited ability to apply the new capabilities in all situations. The process of within-stage development, or horizontal decalage, allows an individual to expand his or her capacity to use the highest level of development in more content areas (Widick, Knefelkamp, & Parker, 1980). Most theorists argue that higher points in a developmental sequence allow more adequate interpretations of an event because they require less distortion of an event or idea and provide for more complex processing.

The Perry Scheme

Perry (1970) borrowed a biological definition that views development "as an orderly progress in which more complex forms are created by the differentiation and reintegration of earlier, simple forms" (p. 44). Although Perry originally defined movement along a hierarchical sequence of stages as development, in 1981 he suggested that "perhaps development is all transition and 'stages' are only resting points along the way" (p. 78). Further, Perry (1981) suggested that growth is not truly linear nor recursive, but perhaps is best represented as a helix with an expanding radius, to show that "when we face the same old issues we do so from a different and broader perspective" (p. 97).

Through a series of 464 open-ended, completely unstructured interviews (84 complete four-year sequences) between 1954 and 1967 originally designed simply to describe the four-year experiences of students at Harvard, Perry (1970) and his associates detected what appeared to represent a coherent pattern of development in the manner in which students functioned intellectually. Further analysis revealed a sequence of cognitive structures used by the students to make meaning in their world, or in Perry's words, to "construe the nature and origins of knowledge, of value, and of responsibility" (p. 1, Perry, 1970). As applied specifically to collegiate environments, the scheme describes how college students make sense out of the information and experiences that confront them in college (Battaglini & Schenkat, 1987).

It is important to note that the Perry Scheme traces an evolution of the "forms in which the students construe their experience" (p. 1) and not the content of the individual's attitudes or concerns. These forms characterize the structures and assumptions by which students function intellectually, view knowledge and truth, experience values, regard the role of authorities, understand their role as learners, and comprehend the meaning of responsibility; it addresses the way students understand the world, their identity, and the way they find personal meaning for their role in that world (King, 1978; Knefelkamp & Slepitza, 1976; Rodgers, 1980; Stonewater, Stonewater, & Hadley, 1986). Perry believed development in these "forms of seeing, knowing, and caring" (Perry, 1970, p. ix) to be transcendent over development of content. Essentially, the theory describes how students move from a simple, categorical view of the world to a more relativistic view, which acknowledges a contingent nature of knowledge and values and recognizes the necessity of making personal commitments. Perry's scheme concerns ethical development in the sense of a person's assumptions about values and responsibilities.

Perry (1970) and his associates recognized that an individual may use a variety of forms at any given moment. However, the scheme is based on the assumption that it is possible "to identify a dominant form in which the person is currently interpreting his experience" (Perry, 1970, p. 3). Perry based the validity of the scheme on the extent to which several observers agreed on the developmental stage of each interview.

Overview of the Scheme

Perry (1970) described late-adolescent development in terms of nine positions, or coherent forms of thought. Each position represents a qualitatively different mode of thinking, or structure, for perceiving the nature of knowledge. As described earlier, the positions are hierarchical and sequential, and movement usually occurs as a result of cognitive disequilibrium. The scheme represents a continuum that describes the steps by which students move from a simplistic, categorical view of the world to a realization of the contingent nature of knowledge and values to the formation and affirmation of their own commitments (King, 1978).

Perry (1970) has grouped these positions in several ways. The sequence of cognitive structures preceding Position 5 describe development from a dualistic absolutism to an acceptance of generalized relativism. Perspectives of Position 5, the pivot point of the scheme, influence the belief that knowledge and values are relative, contingent, and contextual. The final four positions describe how individuals orient themselves through personal commitments in what they now see as a relativistic world.

Alternatively, Perry (1970) describes Positions 1-3 as the Modifying of Dualism, during which students modify their dualistic outlook to allow for the simple pluralism called Multiplicity. Positions 4-6 were called the Realizing of Relativism, where diversity of opinion is recognized and accepted and the necessity of commitment is foreseen. Finally, the Evolving of Commitments, Positions 7-9, trace the development of personal commitments in an individual's actual experience.

Others, most notably King (1978) and Perry (1981) himself, have grouped the positions into four underlying clusters of cognitive structures. Dualism, Positions 1 and 2, represents the belief that knowledge is concrete and absolute, that right answers must be learned from authorities. The dualist uses dichotomous and absolute categories (e.g., right vs. wrong) to understand people, knowledge, and values. The stage of Multiplicity, formed by Positions 3 and 4, represents a viewpoint that, although most knowledge is still considered to be absolute, a diversity of opinions and values is considered legitimate in areas where the truth is not yet known. Because no judgments about the truth can be made in these areas, "everyone has a right to his own opinion" (Perry, 1981, p. 80).

The cluster called Relativism is made up by Positions 5 and 6 (although some consider only Position 5 to be the stage of Relativism). Relativistic students believe that knowledge is dependent upon the context, that interpretations require an examination of the sources, evidence, and logical reasoning used to support judgments. In this way, some opinions have greater value than others and there are issues about which even the most knowledgeable will disagree. With Position 6 comes the awareness of a need to make choices, with the recognition that there may be no adequate way to determine the correctness of the choice.

Finally, Commitment is characterized by Positions 7, 8, and 9. A Commitment is an affirmation or choice made in the awareness of relativism. That is, it is often difficult to determine the correct choice in such matters as career, relationships, and religion; but once a decision is made, the individual must commit to it. Individuals commit themselves to opinions, ideologies, interests, and values with which they will identify (Kurfiss, 1983). Perry (1970) called this a "period of responsibility" (p. 205), in which individuals affirm themselves and their responsibilities in a pluralistic world as they create their identities.

Positions of Intellectual and Ethical Development

The Perry Scheme of ethical and intellectual development includes nine positions, transitions between those positions, and three deflections from growth. Except where specifically noted, the descriptions of these elements of the scheme that follow were adapted from Perry (1970, 1981).

Position 1: Basic Duality

At the first position, a student sees the world as a dichotomy. Generally this dichotomy manifests itself as a strict distinction between right and wrong. Because authorities know the absolute truths, which the student believes to exist in all academic areas, interpretation is unnecessary. Knowledge consists primarily of facts to be learned from the authorities through memorization and hard work. Ability and success are measured by external assessments (e.g., tests, authorities). Morality is viewed from a we-they and good-bad perspective; that is, students learn acceptable behavior by obeying the authorities. Perry (1981) noted that no freshman whom they interviewed still spoke from this perspective, but several saw themselves as having come to college with this view.

Position 2: Multiplicity Pre-legitimate

At the second position students begin to perceive differences of opinion and uncertainty, even among authorities. These lapses from the truth are considered to be the ruminations of unqualified authorities or exercises assigned by qualified authorities so that students might learn to find the answers themselves. There still exist qualified authorities who have the answers; others are either pretenders or are simply wrong. Differences of opinion are considered to be only temporary occurrences as students strive to learn the truth. Such confusion is perceived as resolvable and even valuable as students make their journey to the truth. For example, authorities now are seen to pose problems rather than simply feed students the answers.

The transition phase to the next position is represented primarily by an unconscious new freedom of thought. That is, students are essentially unaware that the exercises they have been given to find an answer or to work out problems has provided them new freedom to think about the issues. Multiplicity has been assimilated into the structures of the early dualistic positions.

Position 3: Multiplicity Subordinate

Diversity and uncertainty become acceptable, but are only accepted as temporary conditions in areas where the answer is still unknown to the authorities. The truth exists, but has not been found yet. More of an emphasis begins for the way to learn over what to learn.

The transition into multiplicity requires students to overcome their impatience with uncertainty, which is now seen as acceptable, even from the authorities who may disagree on some issues. Students at this position begin to concede that there are indeed unresolvable problems and unknowable truths--but these absolute truths still exist. This uncertainty, however, leads to the possibility that any answer may be as good as any other in these unknown areas. The transition phase to the next position therefore reflects an increasing acceptance of uncertainty as a real situation as opposed to a situation manufactured by authorities for teaching purposes.

Position 4: Multiplicity Correlate or Relativism Subordinate

Following Position 3 there are two possible paths for a student to take. One possible route proceeds through Multiplicity Correlate, where everyone is entitled to an equally acceptable opinion. That is, no one is considered wrong. The other route goes through the Relativism Subordinate position, in which a contextual relevance is recognized. That is, some opinions are more valuable than others because of the reasoned support provided for them. Although in 1970, Perry described these positions as alternative developmental routes, in 1981 he expanded the scheme to include the possibility that a path might follow sequentially from Multiplicity Correlate to Relativism Subordinate to Position 5.

Position 4a: Multiplicity Correlate. Perry later called this position "Multiplicity (Diversity and Uncertainty) Coordinate" (1981, p. 84). While looking at the world through the filter of this essentially dualistic position, students still believe that authorities may know the answers in some areas; but where ambiguity exists, any opinion must be considered acceptable. Through this personalistic diversity (i.e., multiplicity), opinions are not judged by any evidence, experience, or context, but are simply related to the person who holds them.

Position 4b: Relativism Subordinate. The student becomes aware of the distinction between an unconsidered opinion and a considered judgment. Comparisons among opinions must be based on more than one factor and the thought processes used in deriving the opinions must be analyzed. Thus, students are now able to think about thinking. Finally, their relationship with authorities has changed from "what they want" to "the way they want you to think" (Perry, 1981, p. 87). Students are better able to distinguish between concrete truths and more complex opinions.

Position 5: Relativism Correlate, Competing, or Diffuse

All knowledge and values are perceived as contextual and relativistic, thereby relegating dualistic issues and absolutes to special cases of complex thought patterns. Complexity is assumed as the general condition; simplicity must be discovered if it happens to be there. From a relativistic viewpoint, not all opinions are considered equally valid. Analysis of context, sources of evidence, assumptions, logic, and inferences will show some interpretations to be better than others, which even may be found worthless. Yet in some areas, disagreement among well-reasoned opinions will still exist. Theories are recognized as models or metaphors with which to interpret information rather than as truth. Authorities base their opinions not on "truths" but on greater experience and expertise than their students. Knowledge has moved from quantitative to qualitative and dependent upon context.

Position 6: Commitment Foreseen

The transition from the previous stage involves uncertainty regarding the ability to determine the correct choice. Students realize that personal commitment is necessary to help them navigate through the relativistic world. That is, students will have to trust their own decisions and recognize that no one will be there to tell them whether they made the right choice. Commitments are those activities that are explicitly important to students, as opposed to those which are taken for granted. That is, commitments represent activities and choices in which a person has decided to invest energy and concern. Perry further describes commitments as those "truths, relationships, purposes, activities, and cares, in all their contexts" that one affirms as one's own (1970, p. 136). From these commitments, individuals then make choices and perform actions.

Positions 7-9: Evolving Commitments

Perry (1970) originally described these positions as "Position 7: Initial Commitment," "Position 8: Orientation in Implications of Commitment," and "Position 9: Developing Commitment(s)," but nevertheless grouped them together in discussion. Essentially, the positions describe making an initial commitment, balancing several commitments, and then affirming the commitments already made. This process, where students "must choose, at their own risk, among disparate systems of navigation" (Perry, 1981, p. 94) occurs again and again throughout life. Finally, ability and success are measured from within the individual. Position 7 describes a state in which students have defined themselves in a particular area of their life. Position 8 describes a level in which stylistic issues have emerged. Position 9 describes a maturity in which persons have defined themselves through both the content of their commitments and the style with which they live them.

From this point on, no major structural changes occur in patterns of thought. The assumption is made that "man's knowing and valuing are relative in time and circumstance, and that in such a world the individual is faced with the responsibility for choice and affirmation" (Perry, 1970, p. 153). Development centers now on the theme of responsibility.

Alternatives to Growth

Although most stage or sequence theories of development do not allow for anything but forward progress (King, 1978), Perry (1970) and associates have identified three possibilities for deflections from growth at critical points of development. During such deflections, a person may "suspend, nullify, or even reverse the process of growth" (Perry, 1970, p. 177). Specifically, individuals may temporize, escape, or retreat when they feel unprepared or overwhelmed by the demands of development.

Temporizing. During a period of temporizing, which is defined as remaining at any stage for over a year, several things may occur. One possibility is that the student recognizes new patterns of thought that lie ahead, but explicitly hesitates in taking that next step. Perry described this as if the student was "waiting or gathering his forces" (Perry, 1970, p. 177). Another possibility is that the student is taking time for "lateral growth" (Perry, 1970, p. 178), where students spread out and consolidate their recently attained positions into other areas of their lives (horizontal decalage).

Retreat. Perry (1970) defined retreat as an entrenchment in, or return to the dualistic orientation of Position 2 and Position 3. Retreat occurs as a reaction to, or an avoidance of, the increasing complexities and ambivalence of the more advanced positions. Perry (1981) later concedes that retreat may have been better described as regression at any stage when the environment becomes too complex. Regression would then describe a retreat into positions already familiar to the student, perhaps to find security and strength to cope with the newly challenging environment (King, 1978). Although regression seems contradictory to the notion of forward progress inherent to stage theories, it may be adaptive and is often limited to a particular situation, thereby serving as what some have called functional regression (Kurfiss, 1983).

Escape. Perry (1970) describes escape as alienation and abandonment of responsibility in order to avoid commitment. Multiplicity and Relativism often provide the means to this detachment by the nature of the structures themselves. Escape as dissociation, or drifting, refers to the "passive delegation of all responsibility to fate" (Perry, 1970, p. 191).

Limitations of the Theory and its Development

Perry (1970) himself noted several major limitations of the development of the scheme. First, the participants in the study were student volunteers in a single college from 1954 to 1967. Second, the developmental scheme was abstracted from oral reports given by the students during annual interviews conducted by the investigators. Third, the validity of the scheme was judged based on the data from which the scheme itself was derived; that is, the study demonstrates "only the coherence between our scheme and our own data" (Perry, 1970, p. 14).

Others have found additional limitations of Perry's scheme, particularly in regard to two key areas: gender and a confusion of underlying constructs. Perry (1970) and his associates derived their scheme of development primarily from a sample of undergraduate men. There were women included in some stages of the research and the group discussed possible differences between men and women, but with few exceptions "the illustrations and validation in this study will draw on the reports of men" (Perry, 1970, p. 16). Further, despite Perry's (1981) claim that the "scheme appears to be a constant phenomenon of a pluralistic culture" (p. 98), there is little evidence of cultural diversity in the sample in any of Perry's

(1968, 1970, 1981) reports. At this point, however, little research has been completed by others to confirm or deny any cultural differences.

The second primary criticism of the Perry Scheme has been with the internal structure of the theory itself. In particular, many researchers have suggested that Perry and his colleagues identified not a single unified theory of intellectual and ethical development, but a combination of theories that confuse cognitive and affective dimensions of development. Most argue that the first half of the scheme (through Position 5) is indeed focused upon epistemological and intellectual development; however, the second half of the scheme (above Position 5) represents a shift into psychosocial development, particularly in the areas of moral, ethical, and identity development based on a Position 5 relativistic perspective (Baxter Magolda & Porterfield, 1988; King, 1978; Moore & Hunter, 1993; Rodgers, 1980). Even Perry (1970) saw Position 5 as a pivot point. King (1978) suggested that finding both issues addressed in one theory has made it more appealing, but that it also has made research difficult. As will be discussed later, most instruments that have been developed to measure the Perry Scheme ignore all but the first five positions.

A third, less important, limitation of the study is the small range of ages of participants. This is a factor because with the exception of Perry's (1970) original study, no investigators have found undergraduates, or even first-year graduate students, who consistently use relativistic thinking (Kitchener, 1982, and current literature review). In some studies, however, this may have been a result of the choice of raters, some of whom were graduate students who may have only just reached the relativistic level themselves. Nevertheless, as King (1978) suggested almost 20 years ago, it still cannot be determined whether there actually are few students in the Committed positions or if the measurement techniques are not successfully eliciting Committed responses. Recent research by Baxter Magolda (1995) has suggested that beyond the undergraduate years, individuals do indeed begin to show a developmental level representative of relativistic, but not necessarily Committed, thinking.

Although no one has criticized Perry specifically, it is interesting that only Perry has seen such a high percentage of his sample show relativistic thinking (and beyond). It is possible that he did not actually observe positions of Commitment as developmental stages. Indeed, Perry has written that the Commitment positions (Position 7 through 9) represent "more qualitative than structural" development (p. 153). Perhaps what he witnessed was some other manifestation of prerelativistic thinking, or as critics have suggested, some form of psychosocial development confused with intellectual development.

Further Theoretical Development of the Scheme

More than any other reasons, the first two limitations described above have led others to attempt to refine Perry's theory. In particular, three sets of researchers have developed what may be considered refinements of the Perry Scheme: Kitchener and King's (1994) Reflective Judgment Model, Belenky, Clinchy, Goldberger, and Tarule's (1986) Women's Ways of Knowing, and Baxter Magolda's (1992b) Ways of Knowing. In their own ways, these theoretical refinements have represented an evolution of the Perry Scheme, although not necessarily developed to be refinements of the Perry Scheme.

Reflective Judgment Model. Although the Reflective Judgment (RJ) Model is not based only on Perry's scheme, several authors have suggested that the two theories appear relatively similar (e.g., Baxter Magolda, 1990; Baxter Magolda & Porterfield, 1988; Rodgers, 1989; Walsh & Betz, 1990). Despite the differences, the similarities between RJ and the Perry Scheme argue for its inclusion here.

The RJ Model has attempted to disentangle development of assumptions about knowledge from those of ethics and identity (King & Kitchener, 1985). Reflective thinking is required when a real problem exists or when a solution cannot be known with certainty. "Reflective judgments are based on the evaluation and integration of existing data and theory into a solution about the problem at hand, a solution that can be rationally defended as most plausible or reasonable, taking into account the sets of conditions under which the problem is being solved" (King & Kitchener, 1994, p. 8). The RJ model describes a developmental progression, based on the study of both men and women, that occurs in the ways that people understand the process of knowing and in the ways they justify their opinions about problems. As individuals develop, they become better able to evaluate knowledge claims and to defend their own points of view because the way people justify their beliefs is related to their assumptions about knowledge (King & Kitchener, 1985, 1994). In sum, RJ focuses on how people reason and arrive at decisions, how they consider the nature and role of evidence in their arguments, how they analyze and synthesize the evidence, and how expert opinion is valued when they make judgments (King, 1978).

Women's Ways of Knowing. Belenky, Clinchy, Goldberger, and Tarule (1986) developed a model of cognitive development based exclusively on women. Their primary concern was that the major theories of human development (as well as most widely accepted conceptions of knowledge and truth have been) had been shaped by men about men. Often, studies of women examine ways that women are like or not like men; consequently, Belenky et al. chose to study only women in order to discover themes that might be more prominent among women. Therefore, when the interview data from the 135 women, 90 of whom were in college, failed to fit neatly into Perry's scheme, Belenky et al. built on Perry's scheme by grouping the women's perspectives on knowing into five epistemological categories: (a) silence, (b) received knowledge, (c) subjective knowledge, (d) procedural knowledge, and (e) constructed knowledge.

Although Belenky et al. (1986) claim that the model does not represent a sequential and hierarchical series of stages, Baxter Magolda (1988) has suggested that the categories appear to increase in complexity when viewed in sequence. Regardless, the categories of the Belenky et al. model seem to share both cognitive structures and related assumptions about knowledge with parallel positions of Perry's scheme (Baxter Magolda, 1990; Rodgers, 1990). However, differences do exist, particularly in the ways in which women view themselves and authorities (Rodgers, 1990). These differences have been called behavioral correlates or stylistic differences, but nevertheless are important because of their implications for teaching and intervention programs. Therefore, Perry and Belenky et al. seem to describe two different styles of knowing within one epistemological structure (Kuk, 1990; Rodgers, 1990). More specifically, women seem not as confident in their abilities in early stages and exhibit more attachment to others in their perspectives of knowledge and learning (Baxter Magolda, 1989).

Ways of Knowing and Reasoning. Baxter Magolda (1992b) has pursued a long history of research into the Perry Scheme of development, particularly in an effort to validate the instrument she developed to measure the Perry Scheme, called the Measure of Epistemological Reflection (e.g., Baxter Magolda, 1987a, 1987b, 1988, 1989, 1990; Baxter Magolda & Porterfield, 1985, 1988; Taylor, 1983). This research has culminated in her adaptation of the Perry Scheme into a model of epistemological development called Ways of Knowing and Reasoning. Baxter Magolda's model was based on research with both men and women and consists of four epistemological levels that appear relatively consistent with the Perry Scheme.

The first three sets of epistemic assumptions, or ways of knowing, were prevalent in collegiate samples: (a) Absolute Knowing, characterized by a view of knowledge as certain; (b) Transitional Knowing, characterized by a growing awareness of uncertainty in some areas of knowledge; and (c) Independent Knowing, characterized by the belief that most knowledge is uncertain. Although the fourth way of knowing, Contextual Knowing, was not observed in many college students, it was borne out by subsequent studies with graduates and graduate students (e.g., Baxter Magolda, 1995). Contextual Knowing is characterized by the creation of knowledge through judgments about evidence in a given context.

Perhaps the most significant element of the Ways of Knowing and Reasoning model is that it identifies gender-related patterns of reasoning structures within the first three levels. The patterns, called impersonal and relational, represent parallel modes of knowing and reasoning that evolve within the same sets of epistemological assumptions, that is, sequence of development. The impersonal pattern is most often exhibited by men and is marked by separation and abstraction; the relational pattern is more often found in women and is characterized by attachment and connection to others. The two patterns, however, seem to converge at the fourth epistemological level.

Gender Issues

Kuk (1990) wrote that "the basic assumption that one could only study White adolescent males and subsequently construct a human development theory based on these studies has created numerous problems" (p. 25). These problems cited by Kuk include (a) a belief that one correct educational model exists for all students, (b) confusion in understanding women's development, and (c) the continuation of a research bias that undermines the validity of research. Indeed, women are often viewed as functioning at lower levels of intellectual development than men (Buczynski, 1993).

Fortunately, researchers have become increasingly conscious of the inadequacies and gaps in developmental theories, especially in regard to women and minorities (Buczynski, 1993; Kurfiss, 1983; Moore & Upcraft, 1990), and have begun to study possible gender and cultural-ethnic differences (Rodgers, 1990). Although psychosocial theories of development have been the focus of many criticisms regarding race, culture, and background, few researchers of the Perry Scheme have studied a diversity in students beyond gender. Students of nontraditional ages have been included in some studies, but for the purpose of validation of the scheme as opposed to exploration of their development.

Knefelkamp, Widick, and Stroad (1976) adapted the Perry Scheme into a model for counseling women by translating the Perry positions to the ways women reason, forming the basis for a counseling process designed to foster movement in women from simplistic to complex thinking. However, Clinchy and Zimmerman (as cited in Baxter Magolda, 1988) and Belenky et al. (1986) questioned the validity of the Perry Scheme in regard to gender. Clinchy and Zimmerman, for example, found that women exhibited a progression similar to the Perry Scheme, but were more hesitant to judge others' opinions and were more insecure with uncertainty. Alishio and Schilling (1984) did not find structural differences between men and women, but differences in patterns of personality and issues of concern.

While some have argued that the Belenky et al. model is substantially different from Perry's scheme (e.g., Buczynski, 1993; Moore & Upcraft, 1990), others have argued for the similarity (Baxter Magolda, 1990). Baxter Magolda (1992b) wrote that although Belenky et al. "found many similarities between Perry's positions and the women they studied, they described perspectives that sounded somewhat different... I found it fascinating that Perry's research on men in the 1950s and 1960s had so much in common with Belenky, Clinchy, Goldberger, and Tarule's research on women in the 1980s" (p. 7-8).

Baxter Magolda and Kitchener and King have used both men and women in their research. Several authors agree, for example, that the Reflective Judgment Model diverges somewhat from the Perry Scheme after RJ stage 3 (Kuk, 1990; Rodgers, 1989). Baxter Magolda (1988) has suggested that the inclusion of women in the construction of these later models accounts for the differences with the Perry Scheme. The Perry Scheme, like most other foundational developmental theories, focuses on development as an autonomous, rational process (Baxter Magolda, 1995). More recent student development theories and refinements of the foundational theories have added a relational dimension (e.g., Baxter Magolda, 1992b).

Research of gender differences and critical examinations of the theories have yielded mixed results. Baxter Magolda (1988, 1990) has indicated that the lack of consistent quantitative evidence of gender differences combined with the existence of qualitative differences suggests that both women and men experience the same range of epistemic assumptions and probably evolve through the same developmental sequence, but possibly differ in the transitions they make and the reasoning approaches they take. For example, Perry's (1970) participants focused on the objective processes for arriving at the truth, while most of the women observed by Belenky et al. (1986) focused on a more interpersonal and subjective approach. Indeed, Baxter Magolda's recent research evolved from the recognition that although men and women enter college with similar cognitive structures, women tend to exhibit a less active role in learning and more attachment to peers (Baxter Magolda, 1990).

An Aside

Earlier, arguments were presented suggesting that the Perry Scheme describes a combination of constructs. Specifically, it has been suggested by many researchers that the positions of Commitment, Positions 7 though 9 (and sometimes 6), describe the ethical development through which individuals might examine their lives and develop their identities. Widick (1977) wrote that "the journey along the nine positions seems to be a perfect operational definition of the examined life" (p. 36). This appears to be a reference to Socrates, to whom Perry (1970) himself refers in distinguishing Perry's description of personal commitments from "habitual never-questioned commitments" (p. 34). However, much of the literature regarding the Perry Scheme never addresses Perry's positions of Commitment, other than to dismiss them as irrelevant. However, had scholars considered these stages more carefully, they may have seen a relevant underlying theme of examination, or even reflection.

Recently, several researchers have begun to argue for the view that the cognitive and affective dimensions of development, which together include knowledge construction, meaning making, and self-awareness, are integrated parts of one process (King & Baxter Magolda, 1996). King and Baxter Magolda (1996) have encouraged that the educational experience should foster not only the cognitive skills needed for critical thinking, but also sense of self, personal maturity, and interpersonal effectiveness. Some specific affective attributes suggested by King and Baxter Magolda are an eagerness to learn, an appreciation of the value of working with diverse others on problems of mutual interest, the desire to make a positive social contribution, and perhaps most noteworthy in the present context, "the will to take personal responsibility for one's views and actions" (p. 163). More to the point, Baxter Magolda (1995) has claimed that the "development of self-identity is necessary for contextual knowing. Balancing one's own and others' perspectives required a self-identity as a foundation" (Baxter Magolda, 1995, p. 214).

Others have also reported similar conclusions. For example, Martin et al. (1994) have suggested that the importance of "the self as an agent who is responsible for his or her epistemic process appears to be an important discovery that emerges in the course of epistemic development" (p. 628). Kitchener and King noted that "the act of taking an intellectual position is a matter of choice for which one is responsible only emerges in the later developmental stages" (as cited in Martin et al., 1994, p. 628).

Baxter Magolda (1992b) reported that both the relational and the impersonal components of knowing are essential to empowering students to construct knowledge and make decisions about career and personal life. Is it possible that the commitments Perry observed were simply manifestations of the relational voice that he had not heard before because he studied mostly men? Whatever the answer, it indeed seems that there is now evidence that Perry was not so obviously wrong as many have believed. Perry does not emphasize a commitment to knowledge or opinion in any of his explanations of the Commitment positions; however, this certainly is a logical extension of the type of personal commitment described in the scheme. Baxter Magolda, King, and others have offered evidence that the construction of knowledge requires such a commitment; indeed, they have even offered evidence (without even realizing it?) that identity is required before one can construct knowledge. Perry's "period of responsibility" may indeed fit within a theory of intellectual development.

Implications of the Perry Scheme

Hoy and Miskel (1991) indicate that the major function of theory is to "describe, explain, and predict regularities in behavior" (p. 3). However, Brown and Barr (1990) have argued that if theory serves only as a description of important variables rather than becoming an active guide for practice, it is of little use. Developmental theories and research must provide useful models through which practitioners can understand students, formulate goals, establish programs and activities, prescribe intervention strategies, and diagnose problems (Brabeck, 1984; Brown & Barr, 1990; Schmidt & Davison, 1983; Widick, 1977). Indeed, Strange (1983) has argued that successful practice requires not only the development of reasonable explanations (theories), but also predictions about what is likely to succeed or fail. Further, the significance of such a plan of action based on theory "does not rest entirely on whether it works exactly as predicted, but that it provides a context for interpreting and responding to actual events as they do happen" (Strange, 1983, p. 3). Therefore, the central question, as posed by Widick (1977), is how well does the Perry Scheme offer guidance in planning and implementing those counseling, consulting, and instructional activities required in practice?

General Implications of Theory

Cross (as cited in King & Kitchener, 1985) stated that "in an applied profession, however, theory and practice must be constantly interactive. Theory without practice is empty, and practice without theory is blind" (as cited in Kitchener & King, 1985). Said another way, paradoxically, "the nature of theory is such that it does not lead directly to practice and the nature of practice is such that it does not proceed without theory" (Parker, 1977, p. 420). More to the point for practitioners, however, Hanson (1990) has asserted that "the strength of the profession will be judged by how well practitioners use information about student growth and development to guide and shape their educational interventions" (p. 270). Indeed, many such interventions have appeared in the literature and the ERIC database (see Table 1).

Table 1
List of Implementations or Suggested Translations of Theory to Practice Found in the Literature and in the ERIC Database

- Academic advising
- Advising
- Anxiety management
- Career counseling
- Career decision making
- Career/life planning
- College orientation
- Conflict management
- Conflict resolution
- Coping with crises
- Counseling
- Critical thinking
- Discipline
- Ethics
- Faculty consultation
- Financial aid services
- Group development
- Interpersonal skills
- Lifelong learning
- Professional development
- Residence hall programs
- Student leadership
- Supervision of staff
- Teaching
- Training
- Training peer counselors
- Understanding alcohol use

Curriculum Design and Teaching in:
  - Agriculture
  - Biology
  - Economics
  - Foreign Language
  - Home Economics
  - Music Education
  - Persuasion/Argumentation
  - Anthropology
  - Composition and Writing
  - Engineering
  - Health Education
  - Library Instruction
  - Psychology
  - Religion
  - Art and Art History
  - Curriculum Design
  - English Literature
  - History
  - Mathematics
  - Public Speaking
  - Teacher Education

Perry (1981) has written that he originally felt an aversion to the application of the scheme, particularly in a prescriptive sense: "Surely educators cannot coerce students into intellectual and ethical development, even if it were ethical to do so" (p. 107). Perhaps not coercion, but certainly understanding college students' intellectual development is at the heart of effective educational practice (Baxter Magolda, 1992b). Indeed, because he considered development to be an interactive process, Perry recognized that practitioners can provide programs and curriculums designed not to prescribe, but to "invite, encourage, challenge, and support students in such development" (Perry, 1981, p. 109). Regarding application of the scheme specifically, Perry (1981) wrote that the scheme "is helpful to the extent that it contributes to the ability of planners and teachers to communicate with students...and to provide differential opportunities for their progress" (p. 107). Indeed, student development practitioners have been challenged to use developmental theories for programming since the early 1970s (Mines, 1985). The Perry Scheme has proved to be a rich source of stimulation for practitioners both in student services and faculty development (Kurfiss, 1983).

Perry (1970) originally limited possible applications to the strict educational procedures of selection, curriculum design, classroom teaching and advising. More specifically, he suggested that using such a developmental scheme in grouping can provide means to identify and support those most in need of assistance, particularly in areas of cultural diversity. Perry indicated that a developmental scheme cannot predict success, but can explain differences among students in perceptions of their instructors. In turn, this recognition of students' differing developmental stages can help an instructor teach them at appropriate levels. King (1978) suggested that the Perry Scheme is helpful in three primary ways: (a) establishing program goals, (b) planning the steps for implementing the program, and (c) evaluating the effectiveness of the program. Also, because individuals at different stages of development apparently learn in different ways and prefer different environments, Perry's scheme and other cognitive developmental theories can help practitioners design interventions appropriate for the differing needs of students. Knefelkamp and Widick (1975) and Widick and Simpson (1978) agreed that the most effective challenges and supports are those that optimally match students at their positions.

Perry (1978) wrote that "many faculties have begged me for a nice, quick, pencil and paper instrument that would measure their success in pushing students along the road to maturity. But I do not like the idea of 'applying' such theories to people, perhaps because I had good old nineteenth-century 'character development' applied to me" (p. 60). Although this sentiment is understandable, many researchers and practitioners have found ways to apply Perry's scheme "descriptively," by inviting them to participate in developmental activities and using appropriate challenges and supports to encourage and assist students through the developmental process. Indeed, recommendations have been made for applying, or translating, Perry's scheme in very nearly every area of campus student life.

METHODS USED TO MEASURE THE PERRY SCHEME

This section will discuss critically how others have attempted to operationalize the relevant construct in Perry's theory. Evidence will be examined concerning the reliability of these currently used measures. The scaling and scoring methods for each will be reviewed. Finally, other alternatives for measuring these relevant constructs in Perry's theory will be explored.

Most definitions of the word theory involve the description of a phenomenon along with the prediction of outcomes. The first section of this paper dealt with the description of the intellectual and ethical development of college students as proposed by William Perry's (1970) scheme. Implicit in Perry's theory, however, is the notion that if one "knows where the student is along the developmental pathway" (Widick, 1977, p. 36), practitioners can predict how they will reason about their current experiences and events. Additionally, practitioners can predict what challenges and supports will be necessary to move them along their path of development. Knowing where the student is, however, requires a valid, reliable, and efficient method by which development can be assessed. Perry's work did not generate such an instrument.

Perry (1968) included in his original report detailed information about the Checklist of Educational Views (CLEV), which was used to screen participants for his first sample; but his attempts to improve it later for those who had requested a quick rating scale proved again to him that "no such things would work" (Perry, 1981, p. 99). Perry (1981) further expresses doubts about creating such an objective rating scheme by discussing the problem of making measurements for an inherently ordinal scale. He suggests that being able to distinguish students in the major levels of development (dualistic, multiplistic, relativistic) might be achieved simply by tuning one's ears to the distinctions among the modes of thought.

There are several methods available with which to measure the Perry Scheme, but no standard assessment procedure (King & Kitchener, 1994; Moore, 1989). Indeed, it is difficult to draw clear conclusions based on much of the research presented in previous sections of this paper because the variety of methods used to assess developmental positions calls into question the comparability of the results (Baxter Magolda, 1987; King, 1978). Consequently, much of the research of the Perry Scheme after the first decade has involved the development of several instruments and techniques.

The assessment problem of Perry's scheme is to determine how far a person has progressed along the sequence of positions identified in the theory. The task is to obtain a representative sample of a person's thinking and then to match the sample with the descriptions of the positions (Rodgers, 1980). Samples of thinking are obtained through various types of instruments that have been developed for measuring student developmental levels according to the Perry Scheme. Although several researchers have developed written objective instruments for the Perry Scheme, the most widely accepted and used methods are semistructured interview techniques and short answer, sentence completion, or essay written instruments. Typically, interview and essay techniques use production tasks, while objective tests use recognition task formats. Some of the more common measurement techniques for the Perry Scheme will be presented after a brief overview of the most important measurement principles used to evaluate them.

Reliability

Reliability essentially is the extent to which instruments provide consistent information. The stability of an instrument is reflected by test-retest reliability, which relies on the similarity of scores for each individual on successive administrations. Equivalence is required when two or more versions, called parallel or alternate forms, of an instrument exist. Internal consistency is a third type of reliability concerned with the homogeneity of items in a test or subtest, or to what extent the items on a test or subtest all measure the same thing. Nunnally (as cited in Moore, 1989) has insisted that Cronbach's coefficient alpha is the single most important measure of internal consistency. Other methods for determining internal consistency include split-half reliability and the Kuder-Richardson formulas.

Lenning (1989) has suggested that instruments must have a much higher reliability, between .80 and .90, for use with individuals, whereas for use with groups a reliability of .60 is sufficient. Similarly, Hanson (1982) recommended that reliability should be over .90 when data from an instrument are to be used for decisions made about individuals, but in the .70 to .80 range when the decisions concern groups. Anastasi (1982) recommended that for the interpretation of individual scores, the standard error of measurement may be much more useful information than reliability. Of course, because the calculation of the standard error of measurement is a function of reliability, higher reliability would be desirable, as Lenning and Hanson have suggested.

Interrater Reliability and Agreement

Since the subjective techniques used to measure the Perry Scheme require the use of trained raters, consistency of the respondents' scores across judges is a primary concern. Two types of evidence are collected to assess this consistency: interrater reliability and interrater agreement. King and Kitchener (1994) have asserted that interrater reliability is the most common index of rater consistency and is typically calculated using a Pearson product moment correlation. Interrater reliability has two problems: (a) the correlation fluctuates dependent upon the heterogeneity of scores within samples and (b) the correlation can be high even when judges do not agree on a single rating (e.g., if one judges always rates respondents one stage higher than the other judge does). Consequently, interrater reliability is often accompanied by a more conservative measure of interrater agreement, which provides an indication of the extent to which the judges assign the same score to a given response. The most common measure of interrater agreement is the proportion of agreement between raters for the dominant position of each respondent.

Unfortunately, most of the validation research described earlier in the paper have used relatively simple measures of interrater reliability and interrater agreement. The simple correlation used for interrater reliability may be less appropriate than, for example, some form of the intraclass correlation coefficient (Shrout & Fleiss, 1979). Similarly, although Stempl (1989) has argued that simple percentage agreement is acceptable as a measure of interrater agreement, Fleiss (1981), Kaid and Wadsworth (1989), Emmert (1989), and others have recommended that more sophisticated measures should be used. Some of these measures of interrater agreement take chance agreement into account, and some also take into account factors like the complexity of the category rating system or the scaling used (i.e., nominal, ordinal, interval, and ratio). Some of these methods are Cohen's kappa (probably the most recommended), Guetzkow's P, Guetzkow's U (measure of disagreement), Scott's pi, and Goodman and Kruskal's lambda. Fleiss (1981) provided a vivid illustration of the differences provided by the various methods (adapted as Table 2).

Table 2

Values of several indices of agreement adapted from Fleiss (1981, p. 234)

Although Stempl (1989) describes percentage agreement as the percentage of items on which judges agree, Fleiss (1981) bases overall proportion of agreement on categories, such that the proportion is calculated as (a+d), based on the following classification table:

		Rater B	
		Given Category (N)	**All Other Categories**
	Category (N)	**a** (.04)	**b** (.06)
Rater A	**All Others**	**c** (.01)	**d** (.89)

	Category Assigned		
Agreement on N	**P**	**N**	**O**
Overall Proportion of Agreement (a+d)	.90	.93	.95
Proportion of Specific Agreement	.94	.53	.80
Goodman & Kruskal's lambda	.88	.06	.60
Agreement on absence of category	.75	.96	.97
Specific Agreement (presence and absence)	.84	.75	.89
Cohen's kappa	.69	.50	.77

Validity

For measurement instruments there are specific guidelines for the determination of validity. Validity can be defined as the "precision with which the test measures" (Ebel & Frisbie, 1986, p. 89) some trait or ability. Specifically, instruments must show three general types of validity, for which evidence may be gathered in a variety of ways. The three types of validity generally recognized as the most important to educational measurement are (a) content validity, (b) construct validity, and (c) criterion validity (Hopkins, Stanley, & Hopkins, 1990; National Committee on Standards, 1966; Wiersma & Jurs, 1985). It should be noted that whereas these issues of validity can be studied empirically for quantitative methods, validity in the qualitative methods depends more on the competence and earnestness of the researcher (Caple, 1991).

Content Validity

Content validity, which is not the same as face validity although some use the terms synonymously, concerns the extent to which an instrument is representative of the issues and concepts relevant to the particular construct of interest (Hopkins, Stanley, & Hopkins, 1990). Content validity is established primarily through the illustration of a logical correspondence between the instrument and the construct (Wiersma & Jurs, 1985). For example, experts in the field of intellectual development can be called upon to analyze items of an instrument for their relevance to the Perry Scheme.

Criterion Validity

Criterion validity concerns the relationship of an instrument to its construct. More specifically, criterion validity relies on the positive correlation of an instrument with another measure of the construct, called the criterion. The criterion measure is often a well-established measurement technique for the same construct. Criterion validity can be analyzed through data collected simultaneously (concurrent validity) or the test can be administered before the criterion is measured (predictive validity). For example, a new objective format assessment for the Perry Scheme may be shown to have criterion validity through comparisons to semistructured interview techniques or other established methods.

Construct Validity

Construct validity is usually defined simply as whether an instrument indeed measures the construct it purports to measure. Like content validity, construct validity is established in part through logical analysis. However, for construct validity, the interest is not in the items themselves, but the relationships among the items. If an instrument purports to measure four underlying dimensions of intellectual development, for example, then items for each dimension should be identifiable. Further, the analysis should demonstrate, usually through factor analysis, that the items work together as expected to measure the underlying dimensions. That is, in simplified terms, the items that were identified to measure Dimension A should be more highly associated with each other than they are with items from other dimensions.

Construct validity can also be shown by providing evidence of convergent validity: that an instrument is related to established constructs to which it is theoretically similar. Conversely, discriminant validity is established through the lack of correlation, or possibly negative correlation, between the measure and constructs that are conceptually different. For example, a measure of cognitive development should be reasonably correlated to measures of moral development (Moore, 1989). Further, component parts of an assessment may be logically related to another construct, such as the absolutist perspective of Perry's scheme or the RJ Model as compared with Rokeach's Dogmatism (Martin et al., 1994). Some have argued construct validity based on the relationships between the dimensions of the construct itself. Sometimes a validation study is used that combines the two into a multitrait-multimethod analysis (Campbell & Fiske, 1959).

Production Tasks

Despite their disadvantages, especially in time and cost, the subjective methods continue to be used more frequently by researchers for one fundamental reason: the production and justification of response. Objective measures typically rely on a recognition format. Most scholars believe that only the production format provides accurate measurement of underlying cognitive structures (e.g., Baxter Magolda & Porterfield, 1988; King, 1990). Basically, a production format instrument uses tasks that are designed so that individuals must spontaneously produce a "stage typical" response based on their repertoire of cognitive reasoning skills, which represents their stage of development (King, 1990; Mines, 1982). Students are usually asked to respond to a series of questions and to justify their responses. The responses are evaluated by judges who compare the similarity of the responses to examples provided in a scoring manual and/or to rating criteria developed for the positions. Further, the qualitative differences inherent in Perry's scheme of development stem from why respondents think the way they do (thus justification) rather than simply what they think (Baxter Magolda & Porterfield, 1988).

Assessment based on production tasks reflects the individual's own approach to the task, as opposed to assessment techniques that require persons to fit their responses into categories defined by the researcher. In other words, respondents are given the freedom to project their own frames of reference into the testing situation. This process is believed to provide the most accurate and complete assessment of their thinking (Baxter Magolda, 1987; King, 1990). Although the most common type of production task is the interview format, some researchers have used incomplete sentence stems and short essays to elicit responses.

King (1990) has outlined several disadvantages of production tasks, which include that: (a) they require trained raters and trained data collectors, (b) they are time-consuming (Mines, 1985, noted that a rule of thumb is 2.5 hours per student) and can be expensive, (c) they require subjective classification of responses, which is subject to bias, and (d) they can be influenced by external factors such as rater fatigue, respondent fatigue, or poor rater training. Although interview formats permit the use of probe techniques to deal with unfocused or confusing responses, they are not conducive to group administration. Although written production task instruments have eliminated the need for trained interviewers, the written production tasks cannot probe an incomplete response. Finally, production instruments present difficult challenges for respondents. Some may feel uncomfortable in interview or essay situations, and some may have difficulty articulating thoughtful justifications under the pressure of the task. These issues, familiarity with a task, difficulty of a task, and the length of a task, may affect performance (King, 1990; Mines, 1986).

Interview Techniques

Perry (1970) utilized a methodology similar to Piaget's observation, collecting data through highly unstructured interviews. In order not to provide any structure for the students' responses, Perry simply asked what stood out for them in their experience of the past year. The development scheme of Perry and his colleagues then emerged from exhaustive qualitative analyses of the ways in which the students seemed to describe their experiences and transformations over their college years (Moore & Hunter, 1993). They used unstructured interviews, which may provide the most detailed information concerning development, but are prohibitively difficult to administer on a scale large enough to validate or to apply the theory. Therefore, later researchers have developed structured or semistructured techniques, both oral and written, to measure the Perry Scheme.

Many early researchers who studied the Perry Scheme believed that the advantages of the interview method outweighed advantages of other data collection methods (Schmidt & Davison, 1983). They believed, as Perry (1970) had, that the only way to understand how people make meaning of their experience is to listen to what they have to say about it. For example, Alishio and Schilling (1984) used interviews that were structured only to focus respondents on the topics of occupation, interpersonal relationships, religion and values, and sexual identity; no other direction was provided. Although interviewing is time consuming and produces information that may not be comparable from one subject to another, the primary advantage is the richness and complexity of the responses provided by the students. The respondent is not required to answer within a limited set of responses as is usually required with objectively scored instruments.

One early attempt to assess students' positions according to the Perry Scheme was an structured interview format that consisted primarily of recognition tasks. This method, developed by Clinchy and Zimmerman (as cited in Baxter Magolda & Porterfield, 1988), required students to choose one statement from a set of statements, which were representative of the Perry positions, and describe the reasons for the choice. The choice of statement was believed to indicate the Perry position of the student; probe questions were used to encourage elaboration, which provided a greater understanding of the student's thinking and also provided additional assessment data. This interview required trained interviewers and trained raters, who analyzed the recorded interviewers for major and minor Perry positions.

Belenky et al. (1986) used an interview and case study approach similar to Perry's original method so that they could hear better what the women in their study had to say in their own words, rather than fit the responses into a preconceived system. The semistructured interviews proceeded at the participant's own pace as she responded to issues including self-image, education, relationships, decision making, personal growth, and visions of the future. The sections of the interview were scored independently by coders who were blind to the various demographic variables collected on each participant. Belenky, et al., first attempted to analyze and classify the data using the Perry Scheme, but they found that the samples of women's thinking they had collected did not fit "so neatly" into Perry's positions. Therefore, they developed their own classification system of women's thinking with which to analyze their data (as described earlier). Using their classification system, Belenky, et al., conducted a contextual analysis of each interview, whereby they reassembled the interviews and reread them many times.

Reflective Judgment Interview

To review briefly, the RJ model describes a sequence of assumptions about knowledge and how those assumptions affect how an individual reasons about problems that are not easily verifiable (Mines, King, Hood, & Wood, 1990). Reflective Judgment is measured through a semistructured interview, called the Reflective Judgment Interview (RJI), that elicits data used to assess an individual's fundamental assumptions about knowledge and how it is gained (King & Kitchener, 1994).

The RJI typically consists of four ill-structured problems that each pose two contradictory points of view that focus on the intellectual aspects of such issues as chemical additives in foods, how humans were created, how the Egyptian pyramids were built, the objectivity of news reporting, and disposal of nuclear waste. Several discipline-specific problems have also been developed. After reading each problem aloud, the interviewer asks a series of standard probe questions to encourage the respondent to produce a response to the problem; further follow-up may be pursued if it is necessary to clarify or focus a response. The standardized follow-up questions of the RJI are designed to tap level of reasoning by asking subjects both to state and to justify their opinions and beliefs concerning the contradictions posed by each dilemma (Kitchener & King, 1985). The probes focus on major concepts of the RJ model, such as reasoning processes used, how opinions are justified, and how alternative interpretations are viewed. For example, to find out about the respondent's assumptions about the certainty of knowledge, an interviewer may ask: "can you ever know for sure that your position on this issue is correct?" (King & Kitchener, 1994, p. 102). While such probing does not guarantee that subjects will use their highest stage of thinking, it does clarify the nature of assumptions currently being used and encourages subjects to take the interview seriously (Kitchener, 1985).

Each interview, which usually takes about one hour (Mines, 1982), is transcribed and then usually evaluated by two certified raters according to the scoring rules (Mines reported that all raters must be certified to ensure that studies across raters will be comparable). Criteria for scoring include assumptions about knowledge, the use of evidence, certainty of knowledge, and the nature of justification (Kitchener, 1985). The scoring manual contains descriptions for each of the RJ model's seven stages, followed by characteristics of each position on three dimensions: (a) cognitive complexity, (b) reasoning style, and (c) openness to diversity. The scores for the responses to each dilemma are assigned independently by the raters and reflect both dominant and subdominant stages.

Mean scores for each subject are derived by averaging these stage scores across all RJI dilemmas and across all raters; these mean scores have been particularly useful for group comparison purposes. Longitudinal change has been examined through several methods, including indication of which stages had losses, which stages had gains, and also changes in the mean score. Kitchener (1985), however, has argued that the individual's mean score represents the best general indication of how that person is likely to reason in similar contexts and also minimizes the effect of any one problem on the assessment.

Reliability and Validity of the RJI. The RJ model generally, and the RJI in particular, have been the focus of a great deal of research (Wood, 1994). King & Kitchener (1994) presented an overview of the findings. Interrater reliability from 24 studies using the RJI has varied from .34 to .97; the median interrater reliability was .78. These reliabilities have varied largely because of the homogeneity of the samples used in the respective studies. Interrater agreement has ranged from 53% to 100% with a median agreement level of 77%. Almost 40% of the studies reported an agreement level of 87% and one-fourth have reported over 90% agreement; only four studies have reported agreement below 70%. Because raters occasionally make errors due to fatigue, a procedure was established to blindly rerate dilemmas for which the raters' independent evaluations are more than one stage discrepant; agreement for these second round ratings has been over 90% (Kitchener, 1985).

Two studies have reported on test-retest reliability using the RJI (King & Kitchener, 1994). One study found a test-retest reliability of .71 for a three-month period with seniors. The second study obtained a test-retest reliability of .87 over a two week period with a larger group consisting of high school, undergraduate, and graduate students. Internal consistency of the RJI has ranged from coefficient alphas of .47 to .99, with a median of .77. Interproblem correlations generally have fallen in the high .40s and problem-total correlations average in the middle .60s. King and Kitchener reported that the four problems are consistent in their level of problem-total correlation, suggesting that each item contributes to the RJI total score.

Because the measurement of reflective judgment requires an interview approach, most research has been limited to small samples; however, there is a large body of evidence that links higher education with increases in reflective judgment (Pascarella & Terenzini, 1991). A large number of cross-sectional studies have indicated that a distinct upward trend exists as a function of educational level (King & Kitchener, 1994; Wood, 1994). RJI scores increase slowly from a high school average of 3.2 to an undergraduate average of 3.8 to a graduate student average of 4.8. Of traditionally aged undergraduates in twenty studies, freshmen averaged 3.6 as compared to an average of 4.0 for seniors; although this change is small, it is meaningful qualitatively. Results from other studies show that college graduates score consistently higher than adults who have not graduated from college, and also show larger increases in Reflective Judgment over those years (King & Kitchener, 1994; Pascarella & Terenzini, 1991). King and Kitchener (1994) reported that several studies have not found significant gender differences, while several others have indeed found differences between men and women. College seniors were found to have significantly higher RJI scores than adult students entering college for the first time (Pascarella & Terenzini, 1991). Kitchener and King (1990) found that both traditional and nontraditional freshman were more similar to each other in RJ stage than either group was to college seniors. Longitudinal studies have also shown a steady, but slow, increase in RJI scores over long time periods. For example, a ten-year longitudinal study showed that 92% of the individuals tested increased during that time, with few regressions. Former high school students had increased an average of 2.9 to 5.5, while college students had increased an average of 3.8 to 5.1 (King & Kitchener, 1994).

King and Kitchener (1994) reported that a number of studies have compared the RJI to instruments that measure other intellectual constructs, including critical thinking, intelligence, verbal reasoning, and formal operations. For example, although students who reasoned at higher stages of RJ also revealed better critical thinking skills, the two critical thinking instruments used in the study were much more highly correlated with each other than either was with the RJI scores (Mines, King, Hood & Wood, 1990). Reflective Judgment has also been compared to identity development, moral development, and psychosocial development, but more for exploratory purposes than to gather evidence for validity of the RJI.

Summary and Conclusions about the RJI. The RJI has generated rich data that have had strong heuristic value for theoretical evolution of the Reflective Judgment Model, as well as the Perry Scheme of development (Mines, 1982). Unfortunately, no studies have been reported that compare the RJI to any other standard measure of the Perry Scheme, such as the Measure of Intellectual Development or the Measure of Epistemological Reflection. Other limitations include the need for certified raters and a fairly cumbersome scoring system (Walsh & Betz, 1990). Mines (1982) argued that the scoring rules underestimate the highest stage of production and do not recognize the complex stage model approach. Averaging major and secondary scores across dilemmas and across raters results in a simple, single stage score. As a result, much of the rich data gained from the interview format is lost. Further, as an interview technique, the RJI is expensive in terms of training, time of administration, and transcription time and costs.

Written Techniques

The interview methodology continues to be used to replicate and extend Perry's work. However, interview techniques are time-consuming and expensive, especially in regard to the training required for both interviewing and rating. Because they are administered individually, interview techniques are particularly prohibitive in outcomes assessment contexts and applied situations (King, 1978). For example, it would not be possible to interview an entire campus or freshman class to determine the best approach for the presentation of a particular workshop. The following sections will discuss the Measure of Intellectual Development and the Measure of Epistemological Reflection, which King and Kitchener (1994) have cited as the most commonly used measures of the Perry Scheme.

Measure of Intellectual Development

The first written instrument used to measure the Perry Scheme was the KneWi developed by Knefelkamp and Widick (as cited in Baxter Magolda & Porterfield, 1988). The KneWi consisted of two essay questions and five sentence completion stems, which sampled the respondents' thinking in several content areas. The rating criteria, based on Perry's original description of the scheme and including structural, attitudinal, behavioral, and language style cues, were used to assign each response to a Perry level. Cross-sectional differences between freshmen and seniors are the only validity data available for this instrument (Baxter Magolda, 1987).

The KneWi evolved into the Measure of Intellectual Development (MID), which is a semistructured production instrument designed to sample respondents' thinking through the use of three essays. The two essays were retained from the KneWi and a third essay was added. Although the MID retains the general stimuli that allow respondents to project their own frames of reference, the three essays were designed to assess how individuals view knowledge and responsibility in three specific content areas: decision making, careers, and classroom learning (Baxter Magolda, 1988). The MID focuses on the intellectual aspects of the Perry model, specifically Positions 1 through 5 and each domain takes approximately fifteen minutes to complete (Mines, 1982).

Responses to the MID essays are coded independently by two trained raters using a rating manual that contains rating cues and examples for each position on a variety of dimensions. Although the rating criteria were originally derived directly from Perry's scheme, they have been elaborated and increasingly standardized based on additional research (Baxter Magolda, 1987). The raters then confer to reach a consensus 3-digit rating for each essay. The scoring system provides stable position ratings as well as two transitional step ratings between each position.

Reliability and Validity of the MID. Reliability studies have correlated researcher ratings of the MID with experts' ratings of the same responses and have provided other interrater reliability data. Absolute interrater agreement on 1785 instruments used in several studies was 54%; the less stringent dominant position agreement among raters, however, increased to 81% (Baxter Magolda, 1988). The reliability correlations reported by researchers (as detailed in Mines, 1982) have been only moderate to high. Correlations between the MID and interview ratings have been reported in two studies at .74 and .77. Correlations of trained raters with expert raters have ranged from .42 to .64 for exact scores and .73 to .87 for dominant positions. Finally, absolute interrater agreement has ranged between .35 and .62 and dominant position agreement has ranged between .74 and 1.00.

Mines (1982) has argued that the focus of the MID on classroom learning and the student-generated responses used in the scoring of the instrument provide it some evidence of face validity (Mines, 1982). Cross-sectional studies by several researchers have reflected the expected differences between freshmen and seniors. Specifically, freshmen have tended to function at Perry's positions 2 and 3 on the MID, while seniors have generally been in transition between positions 3 and 4 (Mines, 1982). Longitudinal studies have been less satisfying (Pascarella & Terenzini, 1991); however, several experimental studies used to examine the effects of developmentally designed classroom activities have shown significant gains in the expected direction (e.g. Knefelkamp & Widick, 1975; Magolda & Porterfield, 1988; Mines, 1982).

Several studies have shown correlations to be in the ranges expected based on the similarities between the MID and other constructs (Magolda & Porterfield, 1988; Mines, 1982). For example, convergent validity has been shown using correlations with conceptual level (r=.51) and ego development (r=.30); unfortunately, evidence based on the Defining Issues Test (DIT) has not been consistent, ranging from correlations of .13 to .45. Although there is some overlap among the similar constructs that have been compared to the MID, the MID seems to measure a reasonably distinct construct. Further evidence of convergent validity has been provided through favorable correlations between the MID responses and ratings based on interviews (Mines, 1982). Mines (1982) also indicated that the MID appears not to be gender-biased, but that cross-cultural effects, especially language, have not been thoroughly examined.

Summary and Conclusions about the MID. As a written response instrument, the MID has provided several advantages over interviews. The primary advantage is that it can be given to many individuals simultaneously, providing a means by which the Perry Scheme can be measured with groups. Although it is a written instrument, it has retained the production format that provides rich, detailed data and has been the standard feature of interview methods used to measure Perry positions. However, unlike most of the subjective methods used early in the research on the Perry Scheme, it uses both a standardized format and a standardized rating system. Although the reliability and validity data have shown some limitations, generally they have shown that the MID can be scored reliably and with acceptable agreement by expert raters. The primary contribution of the MID has been its influence on the validation and development of the Perry Scheme, especially as it relates to the application of developmental activities. Continuing refinement of the rating criteria have yielded an increasingly standardized instrument with higher interrater agreement.

The major limitation of the MID is the extensive training necessary to learn the rating process (Magolda & Porterfield, 1988). Mines (1982) has also indicated that the scoring system underestimates the level of complexity the individual is capable of understanding. Also, the MID does not differentiate among cognitive complexity, cognitive skills, and epistemological assumptions about reality. The content of the measure renders it useful only for collegiate settings (King & Kitchener, 1994). Further, the MID does not account for the stage interaction posited by Perry (1970); rather, the rating criteria are based on narrowly defined content domains that assume a single stage model within each domain (Mines, 1982). Finally, the MID must be scored by expert raters.

Measure of Epistemological Reflection

Baxter Magolda and Porterfield (1988) reported that they had adapted criteria for the development of instruments and manuals from scholars who had researched other areas of development (i.e., Gibbs & Widaman, Loevinger & Wessler). Baxter Magolda and Porterfield cited four criteria they used to design their new instrument: (a) questions must address content areas relevant to the developmental scheme, (b) stimuli, or questions, must separate the content areas, (c) stimuli must probe the respondent's thinking in each content area to elicit adequate data for understanding the respondent's meaning, and (d) stimuli must elicit justification for each response.

These criteria were then used to create the Measure of Epistemological Reflection (MER). The MER is a written instrument that utilizes a production format, standardized stimuli, and standardized rating criteria. The MER consists of a series of questions that address six content domains of thinking and learning, which previous research had shown to be related to epistemological development according to the Perry Scheme: (a) decision making, (b) role of the leaner, (c) role of the instructor, (d) role of peers, (e) role of evaluation in the learning process, and (f) nature of knowledge (Baxter Magolda, 1987). Because separating content domains focuses the respondent's thinking on a particular area, separate series of specific questions were included for each domain.

The general questions, which focus a respondent's thinking on each of the six content areas, require the respondent to produce short answer essays as responses. Separate follow-up questions (usually 3 or 4) encourage elaboration of the reasoning used in response to the general question and also justification for the thinking that led to that particular response. For example, in follow-up probes after the general question about how much students should talk in class, the respondent is asked why that particular level of participation is preferred, advantages and disadvantages to that level or participation, and what type of interaction among members of a class would enhance learning (Baxter Magolda, 1992b). The MER requires about one hour to complete (Baxter Magolda, 1987).

The MER instruments are separated by domain prior to rating. To avoid potential biases, all demographic information is coded and removed from each MER and raters assess all responses for one domain before proceeding to the next set of domain responses. For each content domain, the MER rating manual describes (a) Perry Position 1 through 5, (b) the reasoning structures used within each position, and (c) examples of each reasoning structure. Reasoning structures, which are the justifications the respondents give for their thinking, serve as the unit of analysis in scoring the MER. More specifically, the collective response to the general and specific questions acts as the unit of analysis for the determination of reasoning structure (Baxter Magolda & Porterfield, 1988). The scoring system involves a determination of the modal, or most frequently used, reasoning structure for each domain response (Baxter Magolda, 1985; Baxter Magolda & Porterfield, 1988).

Generally, more than one trained rater evaluates the MER data independently. First, the raters read each response and identify the main reasoning used. Then domain ratings for each response are assigned by matching the reasoning structure evident in the response as closely as possible to criterion responses in the MER rating manual (Baxter Magolda, 1987). The respondent receives an epistemological level rating (Perry Position) and a reasoning structure assignment for each of the six areas (Baxter Magolda, 1989). After all six domains have been rated, the raters discuss discrepancies and determine compromise ratings. Finally, after all domain ratings are assigned, each respondent's MER is reconstructed and a continuous Total Protocol Rating (TPR) is calculated by averaging the respondent's six domain ratings (Baxter Magolda, 1987, 1989; Baxter Magolda & Porterfield, 1988). This interval TPR score serves as the basis for data analysis (Baxter Magolda, 1985). A modal TPR can be calculated that represents the modal reasoning structure across domains, as opposed to the domain rating that represents the modal reasoning within domains. However, Baxter Magolda and Porterfield (1988) suggested that the continuous TPR is preferable because it portrays a more comprehensive picture of the respondent's reasoning.

Reliability and Validity of the MER. Reliability of the MER has been evaluated through interrater reliability, interrater agreement, and internal consistency across domains. Analysis of interrater reliability across several studies (n=752) yielded a correlation of .80; interrater agreement across these studies has ranged from 70% to 80%; internal consistency was .74 (Baxter Magolda, 1987, 1992b).

Naturally, individual studies have provided a range of results. For example, a residence hall study (as reported by Baxter Magolda & Porterfield, 1988) that included 72 men, 93 women, 10 freshmen, 35 sophomores, 69 juniors, 51 seniors, and ages between 18 and 29, determined interrater reliability to be an intraclass correlation of .62. For this same study, internal consistency was calculated with Cronbach's alpha to be .63. Also, exact interrater agreement for dominant ratings ranged from 54% to 68% across domains; agreement within one position ranged from 84% to 96%. Interrater agreement on the TPR scores was 72% for exact agreement and 99% for within-one-position agreement. Seven other studies reported by Baxter Magolda & Porterfield (1988) have shown interrater reliabilities ranging from .59 to .81, internal consistency ranging from .60 to .84, and exact TPR agreement ranging from 46% to 77%. Exact agreement for dominant ratings has been reported variously as ranging between 31% and 83%, while within-one-position agreement has ranged between 47% to 97%.

A further breakdown of interrater agreement demonstrates reliability of expert raters as compared to trained raters. For example, in comparing a variety of training methods, Baxter Magolda (1987) found that the level of exact agreement on TPR scores between experts and raters trained using a workbook-practice-feedback method ranged between 45% and 95%. The seminar approach yielded exact agreement between raters and experts of between 58% and 70%. Intraclass correlations for these approaches ranged between .72 and .97. Baxter Magolda (1985) calculated exact agreement on model TPR scores among experts as ranging between 63% and 73%; however, agreement was as low as 44% between experts and trained raters.

Content validity has been argued for the MER based on the standardized rating manual, which was constructed using an empirical verification process that allows only criteria that have been observed to be included. Adjustment of the manual is made as necessary on the basis of new data. In the initial construction of the rating manual, Perry's scheme (Positions 1 through 5) was used to rate the developmental level of students; specifically, each section provided descriptions and examples of thinking for that domain across Perry positions. During the derivation study, the reasons given by students for their viewpoints were sorted into themes called reasoning structures and added to the manual.

During cross-validation, the structures that were confirmed by the data were retained in the manual and those that were not replicated were omitted. Therefore, overall ratings, which are based on both position and reasoning structures, are based on the empirical data added over time to the manual in addition to initial theoretical descriptions of Perry's model. This process of empirical verification has led to generally increasing reliability of the MER through successive studies. Furthermore, the process has enabled the MER instrument to remain current with the evolving theoretical refinements of the Perry Scheme. Further, because the rating manual was empirically verified using data from both genders, one effect of this empirical adjustment process is that the rating manual contains reasoning structures relevant to both Perry's positions and to the Belenky et al. model (Baxter Magolda, 1988).

The MER has been compared to the MID in at least one study (Baxter Magolda, 1985). In that study, the correlations between the two measures were extremely low (and nonsignificant). However, the MID revealed very little variation across levels of education and indeed, no significant differences were found between the educational levels based on the MID scores as had been the case for the MER. Another attempt to assess criterion validity compared the MER scores for 39 students to ratings from open-ended interviews designed to measure the Perry Scheme. The correlation between the MER and the interview of .93 suggests that the MER measures the underlying structures of the Perry Scheme almost as well as the interview used in the study, which addressed the same domains included in the MER (Baxter

Magolda & Porterfield, 1988). The MER revealed the expected differences as significant between freshmen, seniors, and graduate students. Limitations cited by the authors for the interview study include the small number of participants and the lack of a previously validated interview.

Construct validity data have shown consistently significant differences in cross-sectional studies across levels of education (Baxter Magolda, 1992b). Baxter Magolda (1985) found that group means computed by domain for freshmen, seniors, and graduate students revealed increased cognitive complexity (i.e., Perry Position) for higher educational levels. Further, an analysis of variance showed that TPR scores differed significantly by level of education; post hoc tests showed that freshmen differed from seniors, who also differed from graduate students. These results confirmed that the MER can detect differences expected, on the basis of theory and previous research, between levels of education. Not all studies have been so uniformly corroborative, however. For example, the residence hall study found no significant differences among educational levels, possibly because of the relatively small number of freshmen and sophomores (Baxter Magolda & Porterfield, 1988).

Still, more cross-sectional evidence supports the MER as a valid instrument than not. For example, Baxter Magolda (1987) reported that significant differences were found in four separate samples between freshmen, seniors, and graduate students. Also, one of those studies revealed significant post hoc differences between a group of doctoral and second-year master's students and another group comprised of juniors, seniors, and first-year master's students (Baxter Magolda & Porterfield, 1988). An analysis of data combined from several studies confirmed the differences noted by the individual studies (Baxter Magolda, 1987).

Finally, a longitudinal study had shown significant freshman-to-sophomore gains for both men and women on both the MER and a semistructured interview used for comparison (Pascarella & Terenzini, 1991). Also, a comparison of gender data collected both through the MER and interviews indicated no differences between men and women, providing some evidence that the MER is not biased in favor of either gender (Baxter Magolda, 1988). It should be noted that these data came primarily from students at Position 2 and Position 3. However, there was a consistent pattern of gender differences in reasoning structures evident in both the MER and the interview data, suggesting that although men and women are in the same epistemological position, they may not reason in the same way (Baxter Magolda, 1988).

Summary and Conclusions about the MER. A impressive amount of reliability and validity data, as well as applications using the MER from research to outcomes assessment, have been reported (primarily in Baxter Magolda & Porterfield, 1988). Baxter Magolda (1988) reported that strong evidence had been collected to suggest that the highly structured MER elicited similar data to the more semistructured format of most interview techniques. Specifically, in the 1988 study, no new reasoning structures emerged from the interviews, through the empirical verification process, that were not already included in the MER rating manual.

Presumably, the standardization of the rating manual reduces the amount of inference necessary for rating the MER instrument (Baxter Magolda, 1987). This has been confirmed by generally acceptable interrater reliability and agreement data. Indeed, Baxter Magolda (1987) has suggested that the reliability and the validity of the MER have been sufficiently established to warrant training raters to use the rating manual independently. However, this may have been premature. Although 19 studies were cited by Baxter Magolda for application of the MER (and further reliability validity evidence as a by-product), very little of the evidence has been published. Indeed, eight of the citations were to raw data and four were to thesis proposals or other unpublished papers; only four had been published at that time (information gathered from Baxter Magolda & Porterfield, 1988). No publications based on the unpublished studies or raw data could be found through 1996 in the ERIC database (ERIC on CD-ROM, 1996).

Moreover, a large majority of the reliability and validity studies were performed by Baxter Magolda, with the help of apparently few others. For example, in the primary reference for reliability and validity of the MER (as indicated in Baxter Magolda, 1992b), Baxter Magolda, often assisted by Porterfield, was primarily responsible for the interviews in all studies cited, and served as a rater for every study. While this is not unexpected in the development of an instrument, it is unfortunate that more studies have not been reported by other authors who have used the MER. This is particularly concerning considering the low reliability, especially in the form of agreement, the MER has shown between expert raters and trained raters (see above). Finally, a majority of the reliability and validity studies have been performed at two Midwestern state universities (information gathered from Baxter Magolda & Porterfield, 1988).

Although the MER rating process is not a quick and easy answer to practical assessment, it does rely on a production format and justification of responses that have been the standard of assessment of the Perry Scheme (Baxter Magolda, 1987). It also provides additional specificity of stimuli and specificity of rating criteria over that of the MID (Baxter Magolda, 1985). Despite the fact that written instruments generally result in less detailed information, the MER appears to be a viable means of gathering data about students' ways of knowing (Baxter Magolda, 1992b). Further, unlike the MID, which requires rating by only the distributors, it offers a self-study rater training program for practitioners and researchers interested in using the instrument; the program takes an average of 10 hours to complete (Baxter Magolda, 1987). Indeed, much of the evidence does indicate that the MER can be used in place of interviews to assess epistemological stages fairly accurately, which in turn provides an acceptable alternative for broader sampling than can be performed using interview techniques.

Summary of Production Task Methods

Although the interview techniques elicit the maximum freedom of response and allow respondents to project their frame of reference, they do have certain weaknesses. In particular, Baxter Magolda (1985) has suggested that matching behaviors with underlying developmental concepts is not easy; further, matching those behaviors to the right developmental strands or to the right level of complexity puts the subjective nature of interview techniques in a vulnerable position. That is, the data may be valid, but the interpretations may not be. Consequently, only highly trained raters are used to evaluate interview transcripts. Rest, Cooper, Coder, Masanz, and Anderson (1974) have indicated that varying the test situation and test stimuli, such as in an interview, may have a significant effect on how an individual's responses are scored. In other words, if assessment is not carried out under standardized conditions, then score differences between subjects may be due to different test conditions.

Similar problems exist for the MID, the MER, and other written instruments that must be judged, despite the appearance of standardized rating systems. In addition, the written instruments lack the follow-up questioning possible in interviews, making ratings sometimes even more difficult. For example, the general stimuli of the MID result in a broad range of data, possibly with no justification, making rating difficult; indeed, concern about raters making appropriate inferences prompted the developers of the MID to provide rating services rather than to engage in a widespread formal training program (Baxter Magolda, 1987). On the positive side, studies by Baxter Magolda (1988) have shown that standardized assessment methods (i.e., the MER) are comparable to semistructured methods (i.e., interviews), at least for discerning gender differences. The similarity of the measures provides some evidence that the MER, and possibly other similar written instruments, can perform as well as interviews in collecting rich data.

Generally speaking, however, the rating processes of production tasks make them difficult and time-consuming to score, and potentially costly. Although interviews also require labor-intensive administration procedures, both interviews and the written production task methods require large amount of time to administer. For all production tasks, training is required for raters; for some, training is also required for administration. Also, of particular importance to practitioners, the results of production tasks are not available for immediate feedback (Stonewater, Stonewater, & Hadley, 1986). Further, results are difficult to compare across studies because raters and rating criteria differ.

Recognition Tasks

Because of the disadvantages of the production task format, especially regarding time, cost, and training, several researchers have attempted to create instruments that use recognition tasks. Recognition tasks require respondents to select a response from several options that most closely reflects or describes their own perspective (King, 1990; Mines, 1986; Rodgers, 1980). Instruments with recognition formats can use preference tasks or comprehension tasks (Hanson, 1982). Preference tasks usually use a Likert-type scale through which respondents evaluate a set of stage-typical statements based on a specified criterion, such as their agreement with the item, persuasiveness of the item, or importance of the statement. Comprehension tasks may measure respondents based on how well they are able to paraphrase a statement to demonstrate understanding, how well they match meanings of statements from separate lists, or how well they are able to recall or reproduce a statement at a later time.

An important advantage of recognition tasks is that their standardization of stimuli focuses all respondents on the features of interest in the same way. Similarly, the standardization of scoring reduces the opportunity for rater bias to influence the measurements. Recognition tasks are also generally less difficult than production tasks. The most obvious advantages of recognition tasks are that they are much less time-consuming to administer and to score and the need for less training makes them less expensive. Combined, these factors make recognition tasks more amenable to group administration than production tasks (King, 1990). Larger samples increase the power of statistical tests, thereby increasing the probability of finding small but significant changes or differences (Mines, 1985).

Among the more important disadvantages of the recognition format are that respondents are required to choose from a list of options, none of which may accurately portray the respondents' thinking or preferences about the matter. If respondents are forced to choose the least worst options, the results of the assessment may be only approximate or even inaccurate (King, 1990). Only extensive validity testing can ensure that a recognition test provides appropriate options and otherwise fulfills its expectations. For example, incorrect wording of an item may cause respondents to impose a meaning on a statement that differs from its intended purpose (King, 1990). Also, the objective format does not allow determination of underlying developmental processes used by the student to make the choice. Therefore, this assessment format is most amenable to descriptive evaluations of the stage levels and should be used primarily to classify individuals' developmental levels in a systematic and standardized manner. Therefore, Mines (1985) and others have argued that recognition formats should only be used when the theory has reached a point of stability and typical stage responses have been identified.

Objective Techniques

The need for practical instruments to measure Perry development prompted the construction of objective instruments (Baxter Magolda, 1987; Erwin, 1983). These instruments rely on recognition format tasks which standardize not only the stimuli, as was done with the MER, but also standardize the scoring system. Although standardized rating schemes have been a secondary focus of the development of the MID, the RJI, and the MER, none has overcome the inherent subjectivity of the rating process. Therefore, the following instruments are based on objective scoring processes which require no subjective interpretations of the responses by raters. In general, these instruments have been supported by those who desire to use assessment techniques for practice, but criticized by researchers who supported the production task format.

Paraphrase Technique

Although performed through an interview and therefore more subjective than most recognition formats, Kurfiss (1977) used a paraphrase method based on a recognition comprehension task. The method required the student to read a statement, which represented one of eight Perry positions in one of five topic areas, and then to reformulate the statement in the individual's own words. If the interviewer needed more information, the student was invited to elaborate. The rationale for scoring the responses, based on Guttman scaling, was that the student would most accurately paraphrase statements corresponding to the student's current cognitive developmental level (Baxter Magolda & Porterfield, 1988). In addition to the comprehension task involving sequentially ordered statements, Kurfiss asked participants to react to the statements through Likert scale ratings that assessed agreement with the statement, realism of the item, and how interesting the ideas expressed in the statements were. Finally, after all items from a topic set had been paraphrased and rated, the student ranked the items from most convincing to least and selected the statement most like the student's own views. Based on Perry's stage descriptions, which served as the guide for coding the student's responses, each student was given a score for each topic based on how well the student's answer matched paraphrases in the scoring manual.

Scale of Intellectual Development

The first significant attempt to operationalize Perry's scheme with an objectively scored instrument was by Erwin (1983). The instrument is called the Scale of Intellectual Development (SID). Originally, Erwin (1983) adapted items from an instrument developed by another researcher, the 120-item Scale of Ethical and Intellectual Development (SEID) by Roberts. As a result, Erwin provided little detail about the development of the items beyond that he eliminated, revised, and added items on the basis of their content or relationship to Perry's positions. After keying the items intuitively from his interpretation of Perry's scheme, Erwin administered the 119-item SID to 3,321 entering freshman, of whom 40% were women and the age range was from 16 to 25.

The items on the SID were statements, which were worded both positively and negatively to offset possible response bias, to which the students responded using a four-choice Likert format. After concluding that several of the nine position subscales had reliabilities too low, Erwin (1983) used a factor analysis to determine if a nine position model was appropriate. After removing troublesome items and keying items to their strongest factor, 86 items remained.

Reliability and Validity of the SID. The underlying factor structure of the 86-item SID was interpreted to be comprised of four factors: Dualism, Relativism, Commitment, and Empathy; the reliability for the factors was .81, .70, .76, and .73, respectively. The SID provides a numerical score for each of the four subscales (Stonewater, Stonewater, & Hadley, 1986). Although a person can have a high score on more than one subscale, the scales were constructed such that a high score on one scale would result in low scores on the others (Jones, Newman, Cochran, & Nemec, 1992). The first three factors, Dualism, Relativism, and Commitment, closely followed Perry's categorizations. Erwin's fourth factor, empathy, was not based directly on Perry's scheme but was interpreted from the items that clustered to form the Empathy factor. Erwin correlated the subscale scores to verify that students who scored high on one subscale (presumably the current developmental level) would score lower on the other subscales. Indeed, his expectations were met with the exception of a moderate positive correlation between empathy and commitment.

Whereas Erwin's (1983) first phase explored issues related to the instrument's reliability, content validity, and factorial validity, Phase 2 studied issues related to construct validity, specifically the convergent and discriminant validity of the four subscales of the SID. First, using a subset of 300 students from the original 3,321, the scores from the SID were correlated with two other developmental measures, the Perceived Self Questionnaire (PSQ) and the Erwin Identity Scale (EIS). As expected the Dualism and Relativism subscales were significantly negatively correlated with both the EIS and the PSQ and Empathy exhibited unclear results; however, Commitment was significantly positively correlated to both the EIS and the PSQ maturity scale. While these results certainly provide some evidence of construct validity, it should be noted that most of the correlations were between -.30 and .30; significance may have been an artifact of the large number of subjects in the study (3,321). However, the Commitment subscale indeed did correlate over .40 with several of the EIS and PSQ subscales.

Second, Erwin (1983) compared the SID among groups of all 3,321 students who participated at varying levels in extracurricular activities according to their responses to the Self Descriptive Questionnaire (SDQ). Generally, Erwin found that the greater the participation and responsibility the student reported on the SDQ, the more committed the student scored on the SID. The results of both phases of Erwin's (1983) provide some validity evidence that the SID does indeed measure development of some type. However, because no comparisons were made to an existing measure of the Perry Scheme, the evidence does not validate that the SID measures development along the Perry Scheme (Stonewater, Stonewater, & Hadley, 1986). Jones, Newman, Cochran, and Nemec (1992) also concluded that the SID is not a good measure of the Perry Scheme. In fact, they found that younger students scored higher on SID-Commitment than older students, with a possible curvilinear relationship.

Stonewater, Stonewater, and Hadley (1986), compared SID scores to a paragraph completion task adapted from Allen. Briefly, the Allen Instrument (AI) consists of two essay questions in which students evaluate their educational experience and respond to a situation in which a classmate disagrees with a professor about a point in a biology textbook. Responses to the essays are scored by trained raters who categorize each response relative to Position 2 through Position 5 on the Perry Scheme. One of the four researchers who assisted in the development of the AI had had experience rating responses to the MID essays; all four had extensive experience with the Perry Scheme. Prior interrater agreement for dominant position ratings had ranged between 70% and 80% and test-retest ratings over a one month period had been reported at 86% for dominant ratings. However, in their study, dominant agreement among three raters ranged between 57% and 61%; agreement within one digit of the three-digit rating system were between 63% and 70% (Stonewater et al., 1986).

Stonewater et al. (1986) included in their study 24% graduate students and 76% that were divided approximately equally among the undergraduate years. The sample was selected in this manner to include students who would be likely score at all of the Perry Positions 2 through 5, which is the range of positions usually found in the college population. Stonewater et al. determined that only the SID-Dualism subscale correlated significantly as expected with the ratings based on the Allen measure. That is, the negative correlation confirmed that as the Allen rating increased, the SID-Dualism score decreased. Comparison of the means from analysis of variance results confirmed this conclusion. Jones et al. (1992), based on 290 students with an average age of 20 (but 207 freshmen), were able to establish predictive validity only for the SID-Dualism scale. However, they determined that other important problems existed for the SID-Dualism scale, including the nomological network failure to support construct validity and criterion validity relationships in directions opposite from what was expected. Further, Stonewater et al. found a similar pattern of negative correlations among the Dualism, Relativism, and Commitment subscales as reported by Erwin; but they argued that the negative correlation between the SID-Relativism and SID-Commitment subscales is not consistent with Perry's scheme, as Erwin had suggested.

Summary and Conclusions about the SID. The advantages of the SID lie primarily in its ease of administration and scoring. Unfortunately, the instrument does not have enough validity evidence to warrant its use. The SID-Dualism subscale has potential in the assessment of freshmen, but more research is required for this possibility too (Stonewater, Stonewater, & Hadley, 1986). Erwin himself recommended that future studies might include a more heterogeneous sample of students and might look at student growth as represented by the SID in a follow-up, longitudinal study. Finally, he recommended that studies might also examine the relationship between the SID and structured interviews based on Perry's scheme. Several specific criticisms of the SID include its derivation sample, its validation, and its format.

Although Erwin (1983) concluded that the SID showed promise as an instrument for measuring intellectual development according to the Perry Scheme, he recommended that its use be limited to research, rather than applied, settings until further validation evidence was collected. In particular, a basic flaw in the development of the SID results from using only freshman, 69% of whom were 18 years old, to validate the subscales. Indeed, others have questioned whether freshman can exhibit the range of developmental levels required to validate such an instrument adequately (Baxter Magolda, 1987). Stonewater, Stonewater, & Hadley (1986) have argued that because freshmen usually function below Perry's Position 5, the SID subscales of Relativism, Commitment, and Empathy were structured primarily on the basis of how nonrelativistic students responded to relativistic items. Therefore, because Perry's scheme is built on the assumption that students cannot think at levels beyond their current position, the validity of all SID subscales above Dualism remains unclear.

Other flaws in the SID include that it uses only four categories to represent developmental level rather than a score based on a specific Perry position (Baxter Magolda & Porterfield, 1988; King, 1990). Indeed, one of the four categories, Empathy, does not represent any aspect of the Perry Scheme directly. Although, Erwin (1983) claimed to have examined content validity, it is not apparent that he did more than review the items based on his own interpretations. Moreover, there is no explanation as to how Roberts originally constructed the items. Finally, Erwin did not seek to assess the test-retest reliability of the SID. These criticisms reflect problems that indicate that the SID may not be a valid measure of the underlying construct of Perry's scheme of intellectual development.

Learning Environment Preferences

The Learning Environment Preferences (LEP) instrument was developed by Moore (1989). Moore had been among those researchers using and validating the MID. Not surprisingly, therefore, the LEP is based largely on the MID. The LEP attempts to follow the same approach as the MID in terms of a focus on the features of thinking and learning used by students. Specifically, the LEP assesses student preferences for particular aspects of the classroom learning environment that are associated with increasing complexity according to the Perry Scheme. Based on the arguments that (a) the complex higher positions of Perry's scheme are not representative of intellectual development and are assessed adequately only through interviews and (b) that Position 1 was largely hypothetical in Perry's original study and little empirical evidence has supported its existence in college samples subsequently, the LEP focuses on Position 2 through Position 5 of the Perry Scheme (Moore, 1989). These four positions are assessed across the same five content domains related to epistemology and approaches to learning used in the MID: view of knowledge and course content, role of the instructor, role of the student and peers in the classroom, the classroom atmosphere, and the role of evaluation.

Development of the LEP. The first step in the construction of the LEP was to determine the most frequently used position rating criteria developed over the years for the MID. Over the years that the MID has been used, the rating criteria have been extended and refined. The analysis performed by Moore (1989), then, was based on reviews of actual MID essays collected over several years and rated using these criteria.

The original item pool consisted of 134 statements based on the most frequently used MID rating criteria and excerpts from MID essays that reflected the rating criteria (Moore, 1989). Next, individual items were assigned by two expert MID raters to the specific Perry positions that they were believed to measure; any items which were rated more than one position apart were discarded. Other items that were classified in adjacent positions or were considered transitional or ambiguous were reviewed; if these items could not be clarified, they were also discarded. Through this process, 54 items were removed, leaving 80 items for the first pilot version. In the 80-item instrument, four items represented each position in each of the five domains.

Results of item analyses from a series of pilot tests were used to edit the instrument to a final research version of 60 items. Two item performance indicators were used: correlations among items keyed to the same position and item-total score correlations. Additionally, comments from students regarding wording and interpretability were used in this editing process. Finally, five items, one per content area, were included as meaningless items. These complex-sounding items were taken as quotes directly from actual MID essays. Because they were not based on rating criteria and are not keyed to any position, these items provide a check on whether respondents are choosing preferences because they sound complex (Moore, 1989).

The LEP provides a series of incomplete sentence stems regarding each of the five content domains described above. For example, "In my ideal learning environment, the teacher would..." represents the domain concerning the role of the instructor (King, 1990). The 65 items are distributed across the domains and are rated on a 4-point Likert scale. Each item is rated in terms of its significance to the respondent's ideal learning environment; then for each domain, the respondent ranks the three most significant statements. Patterned after Rest's Defining Issues Test (DIT), all items except the meaningless items in all domains are keyed to a specific Perry position. The major scoring index, the Cognitive Complexity Index (CCI) is derived from a formula based only on responses to the items chosen as most significant by the respondent across all five content domains, producing a score ranging from 200 to 500 corresponding to Perry's Position 2 through Position 5 (Moore, 1989).

Reliability and Validity of the LEP. Moore (1989) used item analyses and two experts to determine the content validity of instrument items as the LEP evolved from 134 to 60 meaningful items. Moore performed a reliability study, using a sample of 725 students from intact groups attending seven institutions, with 47% men and 38% freshmen, 34% sophomores, 10% juniors, and 18% seniors. For each set of items keyed to a specific position across all domains, Cronbach's alpha reliability coefficients were .81, .72, .84, and .84, for Position 2 through Position 5, respectively. Moore concluded, based on further analyses of the items and the patterns of their correlations, that Position 3 lacks the conceptual clarity of the underlying construct apparent in the other subscales. A 1-week test-retest reliability study showed a correlation of .89, suggesting reasonable stability of the instrument over time as well (Saidla, 1990).

A factor analysis was used to provide evidence of construct validity. Moore (1989) determined that four factors comprised the LEP. The patterns of significant loadings on each of the factors led Moore to interpret that the factors did represent the four Perry positions of interest in a general sense. The factor representing Position 2 was clearly defined; however, the factor for Position 3 was less clearly defined and the other two factors seemed to represent two hybrid combinations of Position 4 and Position 5.

The third area of validity addressed by Moore (1989) was criterion-related validity. Moore used two approaches to assess this aspect of validity in the initial development of the LEP. First, Moore used criterion group differences to determine whether the instrument identifies developmental differences among groups that one would expect based on the Perry Scheme. Indeed, based on a sex-balanced subsample of 470 students, the LEP CCI means for class-based subgroups show a steady progression from freshmen to seniors, resulting in a significant difference among the groups. From the converse perspective, no significant differences were expected or found based on sex of the student. Second, the LEP CCI scores were correlated with MID ratings for 215 students at two institutions to test for concurrent validity. The correlation between the LEP and the MID was significant, but only of moderate magnitude at .36. An analysis of variance with groups defined based on their MID ratings, provided a significant difference among groups with a consistent upward progression in the LEP that paralleled the MID ratings (Moore, 1989). Moore concluded that the correlation between the LEP and MID ratings is strong enough to suggest a conceptual overlap between the two instruments, particularly in light of the ANOVA results.

Summary and Conclusions about the LEP. Moore (1989) concluded that the preliminary results were encouraging, but that several cautions must be noted. In particular, although the overall sample size may have been adequate, certain subsamples were drawn too narrowly for the results to be generalizable. For example, only two institutions were used in the concurrent validity analysis described above. Further, the sample was relatively homogeneous and did not reflect an adequate age range (e.g., nontraditional and graduate students) or diversity of students. Also, several types of institutions were not included or under represented.

Although Moore used a factor analysis to provide some evidence of construct validity, he did not attempt to analyze convergent or discriminant validity. However, Saidla (1990) indicated that the LEP was correlated in earlier studies with grade point average (GPA) at .18; one could argue, however, that GPA provides more discriminant validity support than convergent validity.

Moore (1989) indicated that several items did not perform well empirically and may be deleted in future research on the LEP. Also, items representing Positions 4 and 5 need to be clarified so that the underlying factor structure may be strengthened. Indeed, the factor structure shows very strong correlations between factor 4 (Position 3) and both factors 1 and 3 (the hybrids of Positions 4 and 5). These correlations and the strongest correlation, which exists between factors 1 and 3, are not surprising given Moore's data, but do not match the theoretical descriptions described by Perry; Moore indicated that the transition from Position 4 to Position 5 represents the sharpest contrast and transition in the whole scheme. This problem may be a result of the relatively small number of juniors and seniors in the study. More evidence that there may have been insufficient numbers of upperclassmen includes that the means of sophomores and juniors were very similar, indeed, sophomore males had a higher average Perry position than did junior males.

As does the SID, the LEP seems to measure some type of development; however, the LEP has more evidence of content validity, construct validity, and criterion-related validity than does the SID. Indeed, as Moore (1989) indicated, the results of an analysis of variance showed that the mean scores on the LEP progress as expected in relationship to the average MID ratings. The expected progression of average scores across class rank also lend some evidence of validity to the LEP. In particular, Moore suggested that the large difference between freshmen and sophomores is consistent with findings of longitudinal MID studies.

Based on the preliminary reliability and validity data presented in this study, the LEP shows some promise as a reliable and valid objective measure of the Perry Scheme. No trained raters are required to score it, nor is an elaborate rating manual needed. Consequently, the LEP is much less expensive than are production task instruments.

Scale of Adult Intellectual Development

Martin, Silva, Newman, and Thayer (1994) have developed an instrument which is designed to be an objective, written instrument to measure the Reflective Judgment (RJ) Model. The instrument they developed is called the Scale of Adult Intellectual Development (SAID).

Development of the SAID. Items for the SAID were constructed by extracting statements from the seven position descriptions of the RJ model. Items were created for the RJ positions based on 10 dimensions of RJ described by Kitchener and King: role of authority, view of knowledge, use of evidence, understanding of decision making, use of dichotomy, willingness to accept responsibility, complexity of world view, nature of judgment process, attitude toward differing views, and openness to alternative views. Items were not created for certain positions in certain dimensions because of the characteristics of the RJ model. Items consisted of clusters of statements with the hope that the several sentences in each item would reduce ambiguity and increase reliability for each position.

Two judges independently identified the RJ position represented by each item. Any item on which the judges disagreed was revised to more accurately represent the relevant position description. A third judge reviewed the revised instrument. The items were randomly presented to research participants, who responded on a 7-point Likert scale. All items were anchored in the same direction, going from "least like me" to "most like me." Counterbalancing of items was not used because the RJ model is not a binary, polar reasoning system. That is, items reconstructed with negative wording would not necessarily indicate the same position in a negative way; most likely, such items would measure another construct (Martin et al., 1994). This procedure resulted in the formation of 65 items.

Reliability and Validity of the SAID. The factor analysis of items on the SAID-65 resulted in three clear factors that were interpreted to represent three underlying and overlapping epistemological styles: Absolutism, Relativism, and Evaluativism. Subscales were developed from the items which correlated most highly with the factors. Comparison of the items that comprised the subscales to the RJ levels showed that the subscales corresponded closely to the RJ model. All of the Absolutism items came from RJ levels 1 and 2; of 15 Relativism items, nine came from RJ positions 3, 4, and 5; and 16 of the 17 Evaluativism were representative of RJ stages 5, 6, and 7. After the factor analysis, 44 items remained. Intercorrelations among the subscales showed a small negative correlation between Absolutism and each of the others, but a large positive correlation between Relativism and Evaluativism (Martin et al., 1994). Coefficient alpha for each scale was: .79 for Absolutism, .82 for Relativism, and .87 for Evaluativism.

Also in the first study, in which 254 undergraduate students aged 17 to 27 years participated, the SAID-44 was correlated with the SID to show construct validity. Significant positive correlations ($p<.01$) were found between the SID-Dualism subscale and the SAID-Absolutism scale ($r=.40$) and between the SID-Relativism subscale and the SAID-Relativism scale ($r=.22$); however, the SAID

Evaluativism scale did not correlate with any SID subscale. No significant correlations were found between any SAID subscales and the Marlowe-Crowne Social Desirability Scale (Martin et al., 1994).

Martin et al. (1994) performed a second study to further evaluate the validity of the SAID-44. Using 272 undergraduates, the SAID-44 Absolutism subscale was found to be highly correlated (r=.49) with Rokeach's Dogmatism Scale, providing more evidence of convergent validity for the Absolutism subscale; no significant relationships were found between other SAID-44 subscales and the Rokeach scale, including the open-mindedness subscale. The intercorrelations among SAID subscales once again showed a small negative correlation (r=-.16) between Absolutism and Evaluativism and a large positive correlation (r=.65) between Relativism and Evaluativism; no significant correlation was found between Absolutism and Relativism as in Study 1 (Martin et al., 1994).

The third study by Martin et al. (1994) further assessed the reliability and validity of the SAID-44. In particular, test-retest correlations were found to be .75 for Absolutism, .63 for Relativism, and .63 for Evaluativism over a three-week interval; split-half correlations were .57, .53, and .55, respectively. Twelve distractor items were inserted randomly into the SAID-44 to test the validity of the factor structure. Indeed, the distractor items did not consistently load onto any one factor even as the three subscales were replicated reliably. Further construct validity evidence was gathered through the confirmation of the expected relationships with other published scales. In particular, a Locus of Control scale showed SAID-Absolutism to be significantly correlated with the Powerful Others (r=.40) and Chance (r=.30) subscales; the Internal Control subscale was significantly correlated with both SAID-Relativism (r=.28) and SAID-Evaluativism (r=.49). Also, SAID-Evaluativism was positively and significantly correlated with Desirability for Control and Need for Cognition, whereas SAID-Absolutism was negatively and significantly correlated with each. Finally, the Evaluativism subscale of the SAID was significantly correlated with the Marlowe-Crowne scale, although the relationship was small (r=.18).

Summary and Conclusions. Like some of the other instruments described above, the initial validation samples used with the SAID were inherently flawed. Specifically, only students from introductory psychology courses were included. This raises several questions. First, do students who take introductory psychology classes differ in any way from students who do not. Many programs require that a specific program of study begin during freshman year (e.g., engineering, premed) and may not permit their students to take an introductory psychology course. Or perhaps the students are averse to social sciences in preference for arts and humanities or natural science, and therefore simply avoid psychology courses. Second, age and class rank information is not provided for all validation studies. Many introductory courses are taken during freshman and sophomore years, thereby introducing the possibility of a biased sample in favor of underclassmen. Third, all students were enrolled at a single institution, perhaps introducing several confounding factors (e.g., ability, motivation). Finally, no evidence is provided as to the diversity of the sample.

Several other validity issues are of concern for the SAID. In particular, sex differences were found on some of the validation studies. Theoretically, none should have been found, although even some RJI studies have faced this problem. Further, for convergent validity, the SAID was not compared to an instrument for Reflective Judgment (i.e., the RJI) which it purports to measure, but instead used the possibly flawed SID.

Other Techniques

Other instruments have been created with which to measure the Perry Scheme of intellectual development. These techniques share the characteristic of relatively anonymity. In some cases, several authors may have discussed the instruments, but relatively little else is available about them. For example, the Griffith and Chapman's Learning Context Questionnaire is cited by both Belenky et al. (1986) and Moore (1989), but with no information about reliability or validity. They are mentioned here for the sake of completeness.

Foster's Checklist. Baxter Magolda and Porterfield (1988) describe briefly Foster's Checklist, which is an objective instrument that yielded a stage score. Comparisons to a MID essay revealed that subjects scored a stage higher on the checklist than they were rated on the MID. Foster therefore questioned the feasibility of measuring complex constructs with a concise instrument.

Ryan's Instrument. Stonewater et al. (1986) noted that Ryan had designed a seven-item objective-type instrument based on Perry's scheme. The Ryan instrument discriminates Perry position only generally, as dualist versus relativist. The lack of finer distinctions for rating students along the Perry Scheme severely limits the Ryan instrument.

Fago's Scale. Fago (1995) described a 45-item scale he developed similar to Erwin's (1983) SID, but which accounts for less variance than does Erwin's scale. Finally forcing a three-factor solution, Fago found factors that generally paralleled Erwin's. That is, the Dualism factor contained 19 items, 11 of which represented Perry positions 1 and 2; the Relativism factor contained 14 items, of which 9 were from Perry positions 3, 4, or 5; and the Commitment factor contained 12 items, with 10 from positions 6, 7, and 8. Although Fago's scale was developed using only 761 students at a small liberal arts college, it helped to provide evidence that the Perry scheme may be applicable over generations. That is, Fago's research provided similar results over five studies during a 12 year period from 1983-1995.

Parker Cognitive Development Inventory. The Parker Cognitive Development Inventory (PCDI) consists of 144 items that are divided equally among the content areas of education, career, and religion (as described in Baxter Magolda, 1987; Walsh & Betz, 1990; White & Hood, 1989). The items, to which respondents reacted using a 4-point scale ranging from strongly agree to strongly disagree, were initially intended to assess specific positions in the Perry Scheme. However, reliability estimates for several subscales were low and the items were subsequently organized into the three general positional groupings described by Perry (1970): dualism, relativism, and commitment within relativism. Therefore, a score is obtained for each combination of content area and positional grouping, resulting in nine content-positional scores and three composite scores for Dualism, Relativism, and Commitment.

Reliability coefficients of internal consistency for each content and positional rating scale in the range of .81 to .92 (Walsh & Betz, 1990). From a validity perspective, White and Hood (1989) found high positive correlations between the Relativism and Commitment subscales, and each of these subscales was negatively correlated with the Dualism subscale. White and Hood cited further evidence in correlations between the PCDI subscales and other instruments they used to measure Chickering's vectors of development. For example, dualistic students reported that they had made little progress in understanding cultural differences. Some preliminary validity data have revealed differences in development level associated with level of education (Baxter Magolda, 1987). Unfortunately, because little more is available about the PCDI, there is not enough evidence to support the instrument's use as a valid or reliable measure of intellectual development according to the Perry Scheme.

Ways-of-Knowing Instrument. Based on the Belenky et al. (1986) model, Buczynski (1993) created an instrument called the Ways-of-Knowing Instrument (WOKI). By studying the original interviews and responses, Buczynski (1993) created items that were each written to represent one of the five general categories of the Belenky et al. (1986) model. Items then were reviewed by individuals familiar with the Belenky et al. model; any items that were ambiguous or did not adequately represent their respective Belenky stage were either revised or deleted. After this procedure, the questionnaire consisted of 48 items, each of which was to be rated by respondents on a four-point Likert-type scale, ranging from strongly disagree to strongly agree.

The validation of the instrument was based on 348 women at a single public institution. The women's ages ranged from 18 to 25 and 42% were freshman, 10% sophomores, 21% juniors, 22% seniors, and 5% were graduate students. A factor analysis of the data produced five relatively independent factors, one for each category in the Belenky et al. (1986) model. The Silence factor had a coefficient alpha reliability of .69, Received Knowledge had .72, Subjective Knowledge .69, Procedural Knowledge had .80, and Constructed Knowledge had a reliability of .74 (Buczynski, 1993).

Although the instrument and the model were both developed from women's experiences, increasing evidence has suggested that the relational aspects of the model may be useful in describing both women's and men's development (Baxter Magolda, 1995; King & Baxter Magolda, 1996). Given the preliminary evidence reported by Buczynski (1993), the WOKI appears to have relatively strong empirical support that it may measure the categories of the Belenky et al. (1986) model. The factor analysis provided some evidence of construct validity by suggesting five factors that apparently correspond well to the Belenky et al. model. Further, content validity appears to have been satisfied by the use of several persons familiar with the Belenky et al. model to examine the items. Indeed, Buczynski concluded that the validation of the instrument lends support to the validity of the model itself. However, the credentials of these examiners is unknown; further, Buczynski reported that potential participants also examined the items, indicating the possibility that students may have been used as the examiners. Consequently, content validity is not certain. Another limitation is that the sample size may not have been adequate for a factor analysis of 48 items. Another problem arising from the sample is the lack of diversity in the ethnicity and ages of the participants, particularly in regard to minority groups and non-traditionally aged students.

Summary of Recognition Task Methods

Moore (1989) suggested that a reliance on subjective methods is a concern for practitioners who require assessment data on which to base programming efforts and also for researchers who require comparability and confidence of their data across multiple samples. Also, Kitchener (1985) has suggested that student services practitioners need an objective format instrument to provide a practical assessment technique to use in program evaluation. However, practicality means little when accuracy is compromised. The advantage of efficient administration and scoring is lost when a recognition task instrument has questionable validity (Baxter Magolda, 1987). This argument from those who favor more qualitative methods of measuring the Perry Scheme has been fairly well corroborated by validity studies of the most well-known objective instruments. As reported above, each of the objective measures has questionable validity for measuring the underlying structures reflected in the Perry Scheme.

Although objective instruments are easy to administer and score, they also have been subject to the criticism that instruments relying on recognition rather than generation are not appropriate for measuring cognitive structures. In addition, the measures currently available are not connected to specific Perry positions as are the interview or paragraph completion methods. That is, with the production task methods, raters can assign a specific Perry position to the respondent; however, with all of the objective instrument developed so far, only underlying categories can be assigned (e.g., dualism, relativism).

Moore (1989) also criticized several early objective instruments for their "lack of grounding in the ongoing theoretical refinements of the model or in the rich qualitative data collected on the scheme over the past decade" (p. 505). Indeed, significant developments have occurred in relation to the Perry Scheme since those instruments were created. Indeed, even more refinements have been made since Moore's LEP was created--namely, Baxter Magolda's (1992b) book Knowing and Reasoning in College. Certainly, the discovery of two parallel patterns of reasoning has implications for measurement.

Summary of Measurement Section

Several researchers have suggested that the interview techniques and the written format production tasks, especially the MID, were critical to the initial development of the Perry Scheme (e.g., Baxter Magolda & Porterfield, 1988; Moore, 1989). Although these techniques are time-consuming and costly, they provided the rich and detailed data required early in the validation research. Further, the interview format allows the most opportunity both for respondents to project their own frames of reference to a problem and for researchers to clarify ambiguous or incomplete responses (King, 1990). Similarly, others have argued that recognition tasks are not appropriate for theory development or refinement (e.g., King, 1990). Indeed, interviews continue to play a vital role in the assessment of cognitive development (Moore, 1989).

Although subjective measures certainly have been important to the evolution of the Perry Scheme, further refinements may require the use more practical objective instruments. The lack of practical assessment techniques limits the ability of practitioners to link theory to practice; it also affects the ability of researchers to validate the theory across multiple student populations (Baxter Magolda, 1988). More specifically, widespread assessment of diverse populations requires practical assessment techniques; thus, even the most accurate instruments (presumably production tasks) lose their utility for theory refinement because of their lack of practicality (Baxter Magolda, 1987).

The controversy between the advocates of production and the defenders of recognition formats rages on. Studies have shown that individuals can comprehend and prefer an idea or reasoning pattern before they can paraphrase it or spontaneously produce it (Mines, 1985). For example, individuals may be able to prefer an explanation or justification, but not be able to present the arguments by themselves or articulate why they prefer it (King, 1990). Also, the more objective tasks have not shown themselves able to measure the alternatives to growth suggested by the Perry scheme.

Further, Rest (as cited in Hanson, 1982) has argued that individuals tend to recognize and prefer statements at a stage higher than they typically use or are able to understand. The production and recognition formats seem to assess different levels of acquisition of an idea or cognitive skill (Hanson, 1982). However, Moore's (1989) criterion validity comparison of the LEP and the MID showed that the LEP did not consistently overestimate the MID scores as might be expected. At the lower MID scores (e.g., MID of 222), the LEP was indeed higher (e.g., average LEP was 328); but for higher MID scores (e.g., 344-444), the average LEP was under 400 (i.e., 378).

Certainly, the assessment format determines the information obtained. Consequently, an individual may be rated at different stages of development depending on the type of assessment used. It should be noted that most of the researchers who suggest that recognition tasks are less accurate than production tasks base their arguments on the work of Rest, who found such evidence for the DIT and its relationship to Kohlberg's theory of moral development. Researchers have interpreted this to mean that objective measures do not measure accurately the underlying cognitive structures of the theory, but instead measure some related construct (Baxter Magolda & Porterfield, 1988; Kitchener, 1985). Rest, Cooper, Coder, Masanz, and Anderson (1974) boil it down to a question of who does the classification. The production task requires raters to classify the responses based on some standardized scoring criteria, whereas recognition tasks present standardized alternatives representing the scoring categories from which individuals choose, in a sense, their own classification.

Approached from another angle, these conclusions suggest that the recognition task produces a liberal estimate of the person's actual level (Mines, 1986). That is, the score from a recognition instrument may be higher than the person's true cognitive developmental level. On the other hand, the difficulty of a production task may actually make it an underestimate of the person's true level, and therefore make it serve as a conservative estimate (King, 1990). Also, production formats assume that individuals cannot produce a response indicative of a stage higher than their current developmental position, but that they can produce a response lower than their current position (Mines, 1982). A related argument concerns the difficulty of the task involved in the measurement, regardless of format: individuals are likely to score at lower developmental levels on more complex, real-world tasks (Kitchener, 1982).

RECOMMENDED COURSE OF RESEARCH

Accountability issues in student personnel are of increasing interest. This section will describe in detail how a series of studies might be designed that would attempt to validate Perry's theory, especially for this purpose. First, some background on the existing evidence of validity for the Perry Scheme will be presented. Next, a brief overview of the issues of outcomes assessment and accountability will be provided. Then, some of the major difficulties involved in developmental research will be discussed, including assessment problems, measurement issues, selection of sample subjects, and research design difficulties. Finally, the design of a course of research will be explored, including the relevance of qualitative methods to such research, improvements on previous efforts by others, and the possibility of alternative scaling procedures.

Validity of the Perry Scheme

Hoy and Miskel (1991) define theory as "a set of interrelated concepts, assumptions, and generalizations that systematically describes and explains regularities in behavior" (p. 2). Further, the major functions of theory are to predict behavior and to guide additional research in a heuristic sense.

"Theories are useful to the extent that they are internally consistent, generate accurate predictions about events, and help administrators to more easily understand and influence behavior" (Hoy & Miskel, 1991, p. 3). More simply, Moore and Upcraft (1990) suggest that theories should explain phenomenon, predict outcomes, and permit the influence of outcomes. More specifically, the usefulness of a student development theory depends upon to what extent the theory (a) describes the nature of young adult development, and (b) explains the processes of developmental change (Widick, Knefelkamp, & Parker, 1980).

As described above, several authors have provided critical, if not direct, supporting evidence of the validity of the Perry Scheme through validation studies for a variety of measurement techniques. Most importantly, the evidence to be inferred from the work of Baxter Magolda, Kitchener and King, Knefelkamp, Widick, Kurfiss, as well as those who have developed instruments and those who have suggested applications based on the theory, have provided some degree of verification of the Perry Scheme. The research of Belenky et al. (1986) and the most recent work by Baxter Magolda (1992b, 1995) can be considered modified replication studies with similar, but refined, results. Because the validation of developmental theories is a complex process based on indirect evidence, one must carefully integrate results and evidence from many studies. Taken as a whole, then, the findings continue to offer strong evidence for the validity of the Perry Scheme (cf. King, 1978).

Green's Criteria for Valid Theories

Green (1989) provided a set of informal criteria, which include and extend the suggestions of Hoy and Miskel (1991), with which he evaluated developmental theories such as those of Piaget and Kohlberg. Although Green did not use the criteria to examine Perry's theory, his criteria can be used to review the Perry Scheme. The criteria are divided into characteristics of developmental adequacy and scientific worthiness. Developmental adequacy can be examined by looking at (a) temporality, (b) cumulativity, (c) directionality, (d) new modes of organization, and (e) increased capacity for self-control. Scientific worthiness is evaluated by examining (a) testability, (b) external validity, (c) predictive validity, (d) internal consistency, and (e) theoretical economy. Other criteria Green suggests primarily for their aesthetic appeal: (a) texture, (b) novelty, (c) interest, and (d) revolutionary impact.

The opening section of this paper discussed developmental adequacy of the model. In particular, the Perry Scheme is based on the assumptions of hierarchical and invariant sequentiality, or Green's criteria of cumulativity and directionality, respectively. Perry does offer deflections for growth, which have been shown to be realistic, but otherwise maintains the invariant sequences in that everyone must pass through each stage progressively. Temporality is implied in the occurrence of development over time. Each new mode of organization requires the same qualitative change that Perry describes for each of his positions. Finally, increased capacity for self-control implies that as people develop they learn to use feedback so that one's activities can be monitored and adjusted. Perry's positions of commitment fulfill this requirement, as does Baxter Magolda's Contextual Knowing, Belenky et al.'s Constructed Knowledge, and Kitchener and King's Reflective Stage 7.

In the area of scientific worthiness, testability is important so that researchers can verify objectively any claims made by a theory. The theory must be conceptually clear. In order to be testable, a theory must meet two requirements: (a) its constructs must be measurable and the measurement must derive from observable behavior, and (b) its claims must be specific enough to allow predictions to be made. Indeed, Davison, King, Kitchener, and Parker (1980) also indicated that the relationships between the theoretical concepts and observable data must be specified for a theory to be testable. However, Perry's description of the scheme focused on the internal cognitive structures rather than their behavioral or affective manifestations (Widick & Simpson, 1978). Fortunately, others have identified characteristics of the developmental positions, such as that essay tests are hard, uncertainty is stressful, and grading is important (Widick & Simpson, 1978). Also, the development of the rating manuals for the MID and the MER have also provided such behavioral cues.

External validity refers to how accurately a theory describes what we already know about human nature. Predictive validity refers to a theory's ability to foretell new phenomena that are not already known and its ability to generate new knowledge or new research. King (1978), Knefelkamp and Slepitza (1976), and Widick, Knefelkamp, and Parker (1980) have all suggested that the scheme retains face validity because "many people can trace their own intellectual and ethical development through the scheme recognizing themselves, their friends, or their students in the descriptions. Herein lies the explanatory power of the theory and the source of its usefulness in practice" (King, 1978, p. 48). Perry's scheme also has been relatively well accepted for its ability to predict the effects of various factors on students' cognitive development. Widick and Simpson (1978), for example, have used the theory to predict appropriate challenges and supports. Many others have also thought enough of the predictive value of the scheme to base applications on it (e.g., Table 1).

Internal consistency refers not to its psychometric properties, but to the interconnectedness of its assumptions and principles. As has been discussed above, Perry's theory has often been criticized along these lines for its lack of a central theme (but see earlier aside). Further, the deflections from growth can be considered as exceptions which might have jeopardized its internal consistency, but they have been supported empirically (see Mines, 1982). Finally, theoretical economy, akin to parsimony, is an area in which Perry's theory scores relatively high. That is, Perry makes relatively few assumptions in describing an extremely complex phenomenon. The richness and depth of Perry's theory and the attention it has attracted from practitioners in particular provide aesthetic value, as described by Green, to Perry's theory.

Research Evidence of Validity

So through Green's (1989) informal criteria, the Perry Scheme of intellectual and ethical development appears to have solid evidence of theoretical validity. However, the process of research and data collection has also helped establish the theory's validity and has helped scholars to refine the theory where it was potentially flawed (Erwin, 1983). A prime example of this process has occurred in the area of gender. Several, most notably Belenky et al. and Baxter Magolda, have provided evidence, that with a few refinements, the Perry Scheme applies well to both sexes. Other areas of validation, such as the construct validity data reported earlier and the cross-sectional studies of class rank, have been reported earlier. It should be noted that the many successful applications and interventions, for example, developmental instruction (Stephenson & Hunt, 1977; Widick, Knefelkamp, & Parker, 1975; Widick & Simpson, 1978), career development (Knefelkamp & Slepitza, 1976), and others listed in Table 1, provide some evidence of face validity of the Perry scheme (i.e., the theory appears to "work" for many practitioners).

One of the first questions that must be addressed in establishing the validity of any hierarchical model, however, is whether people increase in complexity over time as predicted by the model (King & Kitchener, 1985; Mines, 1986). Mines has outlined three phases to the establishment of such evidence. First, the theoretical and logical consistency of the stages must be demonstrated to have a conceptual internal coherence. For example, Brabeck (1984) indicated that the commonalities among the many descriptions of adult intellectual development may offer some face validity to the notion of sequential change. Kurfiss (1977) used a scalogram analysis of stage comprehension data to determine that the positions reflected a sequence of increasingly complex (harder to paraphrase) concepts, which preference data had indicated were mostly hierarchical.

Second, cross-sectional designs examine whether differences exist between groups that theoretically should be at different stages. For example, seniors are generally more cognitively complex that freshmen, but age may be a mediating factor, especially for older students (Baxter Magolda, 1990). But Cameron (1984) determined that adult learners who were returning to school did not necessarily exhibit higher positions than younger students of the same class rank. The differences observed between groups by cross-sectional studies may be attributable to cohort differences or to differences in the historical influences each group experienced (King & Kitchener, 1985).

Therefore, longitudinal studies, by examining the progression of individuals through the stages, provide evidence that the group differences found in cross-sectional studies are indeed due to developmental factors (King & Kitchener, 1985). Further, Mines (1986) argued that only longitudinal studies can provide evidence of the sequentiality and hierarchical nature of stage theories. Similarly, Brabeck (1984) asserted that testing an assumption of invariant sequentiality requires that the same individuals can be demonstrated to change over time in the predicted direction. Therefore, cross-sectional research designs may suggest the possibility that change occurs, but the actual presence and direction of change must be established through longitudinal data (Mines, 1986). Several longitudinal studies have provided some direct and indirect evidence of the validity of the Perry Scheme (e.g., Baxter Magolda, 1992b, 1995; Kitchener & King, 1990).

Summary of Perry Scheme Validity

From the perspective of college student outcomes, one can know what college students are like when they leave college and still not know how much impact the institution had on the student. Assessment of value-added requires some measure of students in the areas of concern when they enter college because outcomes at graduation are highly dependent upon entering levels (Astin, 1987). Also, longitudinal studies of subjects at the same age who do and do not go to college would help sort out the effects of age and education, which often confound the analyses of development (King & Kitchener, 85).

The most critical area still in need of validation concerns the study of students with diverse personal characteristics (e.g., race, age, socioeconomics) and across multiple and geographically diverse institutions (e.g., urban, rural, large public, small private). Indeed, Stage (1991) recommended that regular measurement of large numbers of college students be undertaken to provide replication and verification of both measures and theories. Additionally, other variables should be studied to ascertain their relationship to student development. For example, several scholars have recommended that certain environmental factors be analyzed for their relationship to intellectual development, such as specific teacher behaviors, evaluation procedures, academic studies, student activities, residence hall programs, peer encounters, faculty interactions with students, intentionally designed learning environments, and faculty perspectives on learning and teaching (Baxter Magolda, 1990; Erwin, 1983; Widick, 1977).

"In terms of populations assessed so far, the Perry Scheme seems established--across sex differences, age groups, different kinds of institutions, and even, to a limited extent, cultures--but efforts to validate the model across large numbers of different populations and settings have been hampered by the assessment instruments available" (Moore, 1989, p. 505). Further, the comparison of various research studies is difficult when the results are based on different measurement techniques (Baxter Magolda, 1987). And although some of the above populations and factors have been studied on a small scale, the most significant impediment to large-scale efforts has been the lack of a valid, practical instrument with which to measure the Perry Scheme (Baxter Magolda & Porterfield, 1988; Moore, 1989; Stage, 1991). Several techniques have been developed for measuring the Perry Scheme, but the objective methods have met with criticism for a variety of reasons (see previous sections). Consequently, the development of an acceptable standardized, objective instrument must be among the first steps of a research project to assess the validity and the value of Perry's scheme for outcomes assessment.

Outcomes Assessment

There is growing interest in outcomes assessment within higher education generally, and student affairs specifically, that will probably continue to increase in the near future (Hanson, 1990; Moore, 1989; Winston & Miller, 1994). As noted by Pascarella and Terenzini (1991), "one of the most basic and persistent questions in educational research has been how one determines the extent to which student change or development can be attributed to the educational experience itself and not to other factors or competing influences" (p. 657). Both internal and external pressures have brought about this increased interest. Externally, legislators, politicians, and private benefactors have become concerned with the cost of higher education and the difficulty in monitoring expenditures within institutions (Moore & Hunter, 1993; Upcraft & Barr, 1990). As a result, states in particular are requiring more accountability of state institutions in the form of outcomes, or more appropriately, value-added assessment programs (Astin, 1987; Erwin, 1991). Further, student development professionals are also faced with issues of accountability and the call to use sophisticated evaluation techniques.

The purpose of such externally motivated assessment is to document students' progress through college and to evaluate the impact of higher education on students. This accountability required by the public is a form of summative evaluation used to determine the overall contribution of higher education to students, which in turn helps the public make decisions about funding higher education and attending particular institutions (Erwin, 1991; Lenning, 1989). Indeed, the NIE Involvement in Learning report (as cited in Erwin, 1991) expressed that higher education must institute systematic programs to assess knowledge, skills, and attitudes from both academic and cocurricular programs; the report emphasized assessment of benefits from student involvement in the campus environment.

While external forces are primarily interested in the summative evaluation of outcomes, most within higher education are interested in both outcomes and processes. Internal pressures to assess outcomes include a desire to gather information for institutional self-improvement (Hanson, 1990). Other internally motivated reasons for evaluation include providing feedback for the continuation, revision, or even termination of programs, such as curriculum or student services (Erwin, 1991). The formative evaluation of processes allows evaluation and improvement of the techniques, time, and tools used to achieve particular outcome goals (Lenning, 1989). Another goal for assessment within higher education is to help students understand themselves better (Lenning, 1989). Rodgers (1980) has argued that without valid assessment techniques, practitioners are, at best, using enlightened intuition when establishing developmental goals, designing appropriate learning activities, planning programs, implementing policies, or structuring environments.

Before any outcomes assessment can occur, educational quality must be defined and goals must be specified (Erwin, 1991; Hanson, 1990). Winston and Miller (1994) reported that many agree that a quality educational collegiate experience included formal academic learning as well as personal development. Unfortunately, goals for students in student affairs programs have historically been vague and difficult to measure. For programs to continue to receive support, student affairs professional must be more effective in communicating to others the important benefits students can gain from their programs and must provide evidence to support their claims (Lenning, 1989). Also, a distinction must be made between the assessment of cognitively-oriented learning objectives (e.g., subject matter knowledge and skills) and developmental objectives, such as critical thinking, ethics, identity, and physical well-being (Erwin, 1991; Hanson, 1990).

Besides a lack of goals, there are other reasons that little outcomes assessment is being done in student affairs. Hanson (1982) in particular has outlined several. First, once theories are developed and go through initial stages of validation and revision, further assessment seems to decline. For example, Hanson noted that less than ten empirical studies were reported in the professional literature from 1970 to 1982 that documented any of the theoretical writings of Perry. Another potential reason is the inadequate training received in research and evaluation at a graduate level. If graduate students are not trained to perform assessment activities, they may not feel comfortable with the assessment process when they are practitioners. Finally, planning in higher education rarely includes the assessment of outcomes. Data collection must be seen as an integral part of program planning and development--after a program starts it is too late to get baseline data. Yet another reason assessments may not be performed is because of competing political priorities. While the student affairs profession subscribes to the developmental perspective, not everyone in higher education does. For example, the Carnegie Commission (as cited in Hanson, 1982) indicated the importance of educating the whole student but emphasized that colleges should foster cognitive development; individuals should take responsibility for development of their social, emotional, and personal skills. Further, Hanson noted that the allocation of resources is usually too small for the often expensive and time-consuming process of developmental assessment.

Perhaps the most important reason is that, despite the growing importance of outcomes assessment, methods to measure student development are insufficiently developed to meet these goals, especially for the evaluation of short-term programs (Hanson, 1982; King, 1978; Mines, 1982). In determining an assessment technique's usefulness, evaluators not only need to consider the reliability and validity of the measure, but also (a) the ease and cost of administration, (b) the ease and cost of scoring,

(c) the ease of interpretation, and (d) the appropriateness of the measure to the concerns of the program (Lenning, 1989; Mines, 1982). Hanson argues further that because there are now so many theories of development, practitioners are likely to use theoretical models for which assessment procedures are readily available. If Hanson is correct, the Perry Scheme would be among the casualties because of the lack of a standardized, objective instrument. Moore (1989) noted, though, that because of the importance of intellectual development as an outcome of college, many higher educational institutions are exploring the Perry Scheme as a major part of their assessment programs.

Difficulties in the Study of Development

The lack of assessment techniques that are sufficiently developed to be useful for outcomes evaluation places the practitioner in an untenable position between the demands for accountability and the existence of few standardized means to achieve it (Mines, 1982). Assessment refers to gathering evidence and therefore requires some type of measurement process. The assessment information is analyzed and used by educators and student affairs practitioners to set goals, to design programs, and to make evaluations or judgments about the value of programs or the best way to improve them (Lenning, 1989; Miller, 1982; Widick, 1977; Winston & Moore, 1991).

Hanson (1990) has identified three primary reasons for the assessment of developmental level. First, the trend to greater interest and involvement in higher education by external constituents has required the establishment of outcomes assessment programs for the purposes of accountability to these external groups. Second, better understanding of the processes of student development and learning will help improve educational practice. Third, assessment data can help students better understand themselves. King (1990) suggested that assessment is needed (a) to monitor educational and developmental progress, (b) to set realistic and informed program goals, (c) to evaluate the effectiveness of programs, and (d) to document an institution's success in fulfilling its mission. Lenning (1980) also mentioned the simple process of description in such a list of assessment purposes. Additionally, Hanson (1982) suggested that the information gained from assessment activities can be applied to program design. Finally, Baxter Magolda (1987) indicated that assessment data are required to refine existing theoretical frameworks.

Developmental Assessment Issues

It seems that one of the greatest difficulties facing student development practitioners in regard to outcomes assessment is the focus on the student. Whereas most student development professionals desire to use a developmental theory such as Perry's to intervene with individual students or small groups of students, the needs of outcomes assessment often require a much broader perspective. That is, many accountability issues concentrate more on issues of overall program success rather than individual student development. So for example, whereas many practitioners prefer production tasks because they appear to measure individual student developmental status better than do more quantitative measures, the latter are more necessary from a broader institutional or program perspective. Student affairs practitioners want detailed information about their students so that they may design interventions and programs to assist in the developmental process. However, the needs of outcomes assessment require information about the change in student development over a specific period of time.

In a sense, the information (e.g., Perry position) required for outcomes assessment is an end in itself, whereas practitioners treat such information as a means to an end (i.e., further development). A willingness to recognize that quantitative measures may measure some aspect of cognitive development according to the Perry scheme with validity may be necessary to further advance this line of research. That is, the awareness that such measures as recognition tasks may be required (and valid) for outcomes assessment, even though the data may not be as rich and detailed as with production tasks, has not been prominent in the Perry literature.

Beyond a fundamental shift in perspective, however, other difficulties face researchers who desire to study outcomes based on cognitive development. Mines (1985) has suggested that assessing development poses additional problems to those typically encountered in behavioral and educational research: the length of time required for developmental stage change and the complexity of stages across content domains. Hanson (1982) has argued further that more must be known about why students develop, specifically what antecedents cause growth.

Issues of Time

Perhaps the most difficult challenge for the design of research in studies of development is the nature of development itself. Several issues of time cause challenge for the study of developmental processes: (a) development is a slow process, (b) the endurance of stable positions is not uniform across individuals, (c) within-stage development differs for each person, some taking longer to consolidate their new structures at the most recently attained position, (d) the traditional college years capture only a small portion of adult development, some of which occurs after college, and (e) cognitive developmental theories express qualitative changes between stages, which imply very little about the time frame or continuity of development.

Development is Slow. It has been reported that most freshmen, and even sophomores, view knowledge as absolute or as only temporarily uncertain (e.g., Baxter Magolda, 1992b; King & Kitchener, 1994; Kitchener, 1982). This would support the report by Mines (1982) that freshmen tend to function at Perry Position 2 or 3 and seniors at Positions 3 and 4, as measured by the MID. Seniors generally seem to view knowledge as absolute in some content areas, but rely more on their own opinions in areas where knowledge may be viewed as uncertain (King & Baxter Magolda, 1996). Seniors may believe that many answers exist for every question and that there is no way to use reason to resolve the disagreements with certainty (Kitchener, 1982). These reports, and others cited in previous sections, confirm that development generally is not great between freshman and senior years of college, perhaps as little as one-half to one position on the Perry Scheme.

The challenges posed by the slow process of development are numerous. From a practice perspective, for example, the duration of most programs and activities are typically much shorter than the expected duration of any developmental stage, making assessment pointless or misleading (Mines, 1985; Winston & Moore, 1991). Kitchener (1982) has suggested that this precludes the possibility that a typical student services program would have significant, measurable impact on any specific developmental domain, because four years or longer may lapse between stage progression. Also, reasoning capacities do not advance as high or as quickly as most educators would like (Kitchener, 1982).

Development Progresses Uniquely among Individuals. King and Baxter Magolda (1996) indicated that the "task of enhancing development remains a daunting one because students' starting points often differ so sharply from the developmental goals that educators envision" (p. 168). Many practitioners and educators erroneously assume that students can reason at probabilistic levels (Kitchener, 1982). Perhaps the most significant implication of King and Baxter Magolda's suggestion is that students differ. Indeed, Kurfiss (1983) has argued that the use of the Perry Scheme in both research and teaching has been hampered by the lack of consistency within individuals that would enable practitioners to make global judgments of an individual's level of functioning.

Mines (1986) has indicated that the length of time required for stage acquisition poses no problem if all intervals are equal. However, because individuals differ greatly in their development, the timing of stage changes presents an important issues for longitudinal research in particular, where repeated measures are typically made at standard intervals for all participants. Also, these regular intervals (e.g., freshman and senior years) may mask changes that occur during the interim (e.g., sophomore and junior years) and are no longer evident at the end of the period (Winston & Moore, 1991). Further, Fischer (as cited in Kitchener, 1982) argued that there may also be a great deal of individuality in how progression from one level to the next occurs. Kitchener (1985) and Winston and Moore (1991) have implied that assessment must also account for why or how this change occurs, because no adequate explanation exists.

Development Occurs Within Stages. The process of development may involve periods of advancement and consolidation; several dissonant experiences may be necessary before thought patterns are reorganized (Kitchener, 1982). Widick (1977) implied that the developmental pattern does not proceed at the same rate in all areas (e.g., career, relationships). Hunt (as cited in Schmidt & Davison, 1983) suggested that development should be considered as both stage progression and stage articulation. Horizontal growth as a goal for intervention would help students to feel comfortable in their new perspective and to extend their primary cognitive structures into more content areas.

Additionally, because most developmental models only describe the qualitatively distinct thought patterns that occur, only epochal development can be measured; fine-grained within-stage differences are difficult to capture with such models (Kitchener, 1982, 1985). Kitchener's (1985) comment that there are subtle within-stage differences that the RJ Model does not capture is equally applicable to the Perry Scheme given the current state of research. Whitt (1991) noted that developmental research examines change over time, a process that cannot be adequately portrayed by measuring an individual only at the beginning and end of some time interval. Mines (1985) has described such fine-grained development as microdevelopmental changes, which specifically are skills and behaviors that represent varying degrees of stage mastery. Microdevelopmental changes are assumed to occur in smaller time periods than global stage changes. Mines (1982) has argued that researchers must begin to look at specific cognitive skills that are part of a given stage and develop instruments to measure them.

Development Occurs after College. Although both cross-sectional and longitudinal research has supported the conclusion that students develop between their matriculation and graduation, it becomes apparent that students are just beginning to construct knowledge as they graduate from college (King & Baxter Magolda, 1996). For example, the conclusion was cited above that many seniors only develop as far as Perry Position 4 during college. Kurfiss (1977) determined that there was little difference between freshmen and juniors in comprehension of stage descriptions and that both freshmen and juniors preferred Position 4 statements to those of other positions. Kitchener (1982) suggested that fully mature cognitive structures are unlikely to develop until the middle to late twenties. Further, King and Baxter Magolda cite that Kegan has estimated that only half of adults have achieved an internally generated sense of self. Therefore, because development does not end when a student graduates from college, only a specific segment of the developmental process can be measured (Mines, 1985).

Development is Qualitative. Because intellectual development represents qualitative changes in the ways in which individuals think and reason, no assumptions are made that the stages represent equal intervals of growth or that the stages are equally distant in some respect (Mines, 1986). A basic premise of stage models is that the stages are discrete reasoning processes and that it is the manner in which the individual reasons about a problem that is of interest. Attempting to use quantitative assessment techniques may present an artificial picture of continuity of the development process, which has a direct bearing on scaling procedures used to study development and to analyze data statistically (Mines, 1986).

Without quantitative techniques, assessment must be limited to nominal or ordinal scaling procedures. Mines gives as examples the use of a Guttman-type scale as a conservative approach to the sequential and hierarchical arrangement of stages or the use of an unfolding model to analyze predominate and adjacent stage usage. Because defining a Guttman scale can be very difficult, however, ordinal scaling offers more flexibility (Mines, 1986). Another possibility is to sacrifice the apparent precision of a single stage score by using a scoring system of multiple scores that maintains the complexity of stages (e.g., a dominant and subordinate stage rating).

Complexity of Stages

Hanson (1982) has suggested that qualitative descriptions should not be used to summarize developmental status because the responses of most students are not consistent enough across content areas to be classified accurately into any one stage. This is a consequence of an increasingly important topic of developmental theory: complex versus simple stages (Davison, King, Kitchener, & Parker, 1980; Kitchener, 1982; Martin et al., 1994; Mines, 1982, 1985). Another difficulty of the complexity of the developmental process is the simultaneous development in a variety of domains (e.g., cognitive, moral, psychosocial). In particular, identifying the behaviors representative of a particular developmental strand presents a challenge to researchers.

Complex versus Simple Stages. The data in the area of social-cognitive development support the assumption that stages are complex, not simple; that is, any individual may use reasoning structures that vary according to the context and the content domain (Davison et al., 1980; Mines, 1982). As Kitchener indicated, because complex stage models offer less precision, "one cannot say that a subject's reasoning is 'in' a particular stage" (p. 86). Further, the data support temporary regression in the developmental process (Mines, 1982). This complexity and unevenness of stages across content areas, which prohibits a global, all-or-none assessment approach, occurs primarily as lower stage thinking is replaced gradually by higher cognitive processes (Davison et al., 1980; Mines, 1985).

Kurfiss (1977) concluded that a student is most likely to advance more quickly in content areas where the student is persistently engaged. But Kurfiss also argued that identification of an individual's current position is compounded by the fact that when entering new and challenging situations, people temporarily tend to use earlier positions (functional regression), causing further within-stage inconsistencies across content domains. Indeed, Kurfiss suggested that the Perry Scheme itself suggests that the "discovery of relativism may first come in academic or social areas and then be applied to the search for personal values and commitments" (p. 569).

Scoring Procedures. The problem of stage complexity is particularly troublesome in the determination of scoring procedures used in assessment techniques. For example, Hanson (1982) and Kitchener (9185) have suggested that assessments should not result in a stage classification but in a probability statement that indicates either the frequency of a response level or the likelihood of using a particular reasoning style when faced with specified problems. Davison et al. (1980) suggested that the predominant, or modal, level best describes an individual's current way of reasoning. Because higher level thinking increases as consolidation occurs at a new stage, Davison et al. argue that the reasoning style observed most often best represents an individuals current stage. Therefore, when a new stage is attained, most reasoning still occurs at the previous stage; but as the stage is consolidated across content domains, it becomes predominant--until it is replaced by the next stage.

Kitchener (1985) has argued that the mean score, used by the RJI, represents the best general indication of how an individual will reason when faced with similar problems. However, others have argued that the highest level attained best represents an individual's current cognitive complexity. A variation of this approach is to use optimal level assessment. Under natural conditions, individuals respond to problems without any practice allowed or instruction as to what type of response is expected. However, under optimal conditions, subjects are given the task beforehand and provided instruction concerning better ways to perform the task. An example of this procedure is the Prototypic Reflective Judgment Interview, which allows respondents an opportunity to think about the topics in advance. This procedure is designed to elicit the highest stage of thinking that the person is capable of producing (King & Kitchener, 1994). In addition to probabilistic, modal, highest, and average levels, Mines (1985) listed scoring schemes based on percentages, the use of cutting scores based on cumulative distributions of responses typical of each stage, and strong scalogram analyses.

At worst, using the modal, highest, or average position assumes a simple stage model of development; at the very best, these scoring methods undervalue the complexity of stage information obtained from the assessment (Mines, 1985). Moreover, because other stage information is lost (such as horizontal decalage), the use of the highest stage as a score suggests that an individual should always produce responses typical of their score (Mines, 1985). However, people are not always able or motivated to produce to their highest abilities; in fact, a person may exhibit lower stages in all responses but one. Mines has suggested that it may be interesting to know only the lowest level of response, which may represent the stage typical for the individual when under stress. The modal score underestimates the highest stage of production or comprehension and provides no indication of the prevalence of more higher of lower responses. Mean stage scores account for higher and lower stage responses but still ignore stage variation. Mean scores also result in a conservative estimate of stage level, especially when averaged across content domains or across raters (Mines, 1985).

In the scalogram technique a task or skill is selected that an individual at a given stage should be able to complete successfully, but an individual at the stage or level below should not be able to complete. This procedure predicts a sequence of steps in acquiring developmental skills within a specific content domain and is scored using an adaptation of a Guttman scaling procedure. Although the scalogram method eliminates several problems inherent in the other methods, it is constrained to a single content domain (Wise, 1986). Scalogram analysis is used to order a group of tasks into a linear hierarchy and to evaluate the extent to which each task is prerequisite to the tasks above it in the hierarchy. However, scalogram analysis is limited to linear hierarchies and results in measurement based on single items, which often results in low reliability (Wise, 1986).

Reporting the percentage of responses at the highest stage recognizes more complexity but also puts less emphasis on the lower reasoning structures exhibited, particularly in a case where responses range over three stages. A similar method provides a predominant stage rating plus secondary ratings. Finally, Mines reported that Loevinger has used ogive rules of cumulative distributions as cutting scores. This method does not provide a breakdown of stage responses for each student and therefore does not convey the interplay of different stage assumptions (Mines, 1985).

Strands of Development and Specification of Behaviors. Ego development, moral development, intellectual development, and others all occur simultaneously, making it difficult to separate these strands of developmental phenomena in the interpretation of assessment data (Baxter Magolda & Porterfield, 1988; Kitchener, 1982; Taylor, 1983). The difficulty of matching behavior to the appropriate underlying developmental process, especially when combined with a complex stage perspective, makes assessment complicated. Further, some behaviors from earlier developmental levels are repeated in only slightly different ways in later stages (Baxter Magolda & Porterfield, 1988). All of these factors can lead to inaccurate interpretation of assessment data. Not only are clear and specific descriptions of qualitative differences among stages necessary to distinguish intellectual development from other strands of development, stimuli are needed that elicit these specific differences in sufficient detail (Taylor, 1983). Therefore, behavior patterns indicative of different developmental domains must be identified and isolated (Walsh & Betz, 1990). Mines (1985) also argued the point that specific skills rather than global descriptors must be incorporated for each stage. If patterns of core behaviors and cognitive skills can be identified, it should be possible to better assess students' developmental levels and thereby provide more adequate intervention strategies (Widick, 1977).

Measurement Issues

Moore (1989) has argued that beyond being empirically sound, a measure of the Perry Scheme must "be grounded in the intellectual heritage of, and ongoing research with, the Perry model and heuristic in terms of its ability to foster sound research" (p. 504). Unfortunately, the process of validating empirical soundness is necessarily more than a preface. Without careful consideration of the psychometric properties, the validity of developmental models may never be adequately established (Mines, 1986).

Process or Status

One of the most critical psychometric dilemmas that must be addressed is whether one is "trying to assess a developmental process (that is, a social-cognitive stage), or are we trying to capture a developmental moment as a landmark or sign of a given individual's development" (Mines, 1982, p. 66). Light, Singer, and Willett (1990) described the dilemma as a question of whether to study the status of students at one point in time (e.g., what Perry position they are) or changes in status over a period of time (e.g., how much they gained or improved over the last year). Moreover, the purpose of the assessment may determine the information that must be collected. Researchers and practitioners must consider what depth or richness is required. For example, interviews allow access to such process information as whether a student has retreated or is temporizing. Objective measures simply report the current developmental level of the student.

Mines (1982) noted that when a process phenomenon is reduced to a static description, measurement error is introduced into the assessment. Hence, the problem described above that differences occur in scoring based on production or recognition measurement formats, presumably with more error in the recognition formats that "capture a developmental moment." Hanson (1982), therefore, has argued that because of the importance of underlying processes to cognitive development, an assessment process is required that provides data beyond the developmental status of individuals, including how they change, when they change, and what conditions seem to influence the change. Similarly, Light, Singer, and Willett (1990) suggested that because students come to college with diverse backgrounds and skills, the research focus should be on growth rather than status. Although one might argue that changes in developmental status show growth, whether that is assessed through rich and detailed information or repeated snapshots of developmental levels.

Standardization

The impracticality of complex, subjective processes hinders validation and application of theory because widespread assessment of diverse populations requires practical assessment techniques (Baxter Magolda, 1987). Another consideration is that standardized conditions must be used when comparison are made among individuals. If somewhat different stimuli, probe questions, interviewers, or test formats are used, the responses across respondents may lack any consistency or coherence necessary for comparisons

Although some believe that accuracy must be sacrificed for the cause of practicality, Kohlberg (as cited in Baxter Magolda, 1987) has outlined three phases of developmental research that supports the necessity of both qualities. The first phase of research endeavors to define a development model by identifying and validating the processes of growth. In this phase, unstructured and semistructured assessment techniques using production tasks allow for the freedom of respondents to project their own frames of reference and require the production of responses that provide the full range of responses and processes that define the developmental levels.

Once the developmental scheme has been detailed sufficiently, thereby defining the limits of the scheme in terms of frames of reference and underlying processes, researchers can use this knowledge to construct standard assessment procedures without fear of limiting respondents to an arbitrary set of criteria. When these standardized assessment procedures achieve acceptable reliability and validity, they can be used in the third phase of developmental research. The third phase of research is to refine the theory and requires the practicality (in terms of time and cost) necessary to assess large and diverse populations of individuals.

Some might add a fourth stage to this research process, in which theory is tested in applied situations (e.g., Baxter Magolda & Porterfield, 1988). The process of linking theory to practice also requires practical assessment techniques that can be used by practitioners. For example, Hanson (1982) recommended both computer-assisted assessment and instruments that can be self-scored by students as two means to make the process more practical. Most simply argue for a written objective instrument (e.g., Buczynski, 1993; Erwin, 1983; Moore, 1989).

Social Desirability

Written objective instruments have been suspected of vulnerability to a social approval response set (Borg & Gall, 1983; Grotevant & Adams, 1984; Hopkins, Stanley, & Hopkins, 1990). That is, respondents may try to present themselves in a favorable light, making themselves "look better" by how they respond to certain items. In the study of development, one may try to appear more developed by choosing what appear to be more complex options. Two methods have been used by developmental researchers to determine whether individuals are responding to written instruments in ways that might improve their ratings. One method, based on the Marlowe-Crowne Social Desirability Scale, is used to determine whether individuals are using response patterns that would indicate a bias toward social desirability, or answers that enhance a respondent's score. Adams, Shea, and Fitch (1979) and Grotevant and Adams (1984) used the correlations between the Marlowe-Crowne and subscales of an identity scale to determine that the instrument was not vulnerable to this bias. Martin et al. (1994) also used the Marlowe-Crowne for similar purposes in the development of the SAID. The second method is to include theoretically irrelevant, or meaningless, items that appear complex and therefore attractive (e.g., Moore, 1989).

Precision of Assessment

Assessment data are useful only to the extent that what is measured provides appropriate information for the decisions that must be made using the data (Hanson, 1982). For example, an instrument that only allows gross judgments from one major developmental stage to another would do little to distinguish the modes of thinking within a freshman class, most of whom are known to fall within the dualistic mode of thinking. In general, Hanson suggested, the more likely data will be used to make decisions about individuals, the more accurate and specific the assessment procedure must be. Kitchener (1985) and Mines (1985) have also discussed the need for instruments that can measure more accurately at a "microdevelopmental" level. For example, it may be possible to define stages more precisely by identifying content differences in stage acquisition or identifying component skills representative of each position.

Kitchener (1985) also presented the other side of the issue--the macrodevelopmental argument. Specifically, she asked "do the stage differences make a real difference in how adults understand and operate in the world around them?" (p. 91). Taking the argument a step further, what specificity of knowledge of an individual's or group's developmental status is required to help practitioners design programmatic interventions or learning experiences? Indeed, most of the applications employed using the Perry Scheme have not looked at specific positions, but at the underlying classifications (i.e., dualism, multiplicity, relativism). For example, the classic studies of developmental instruction discussed only how to challenge and support dualists or relativists (e.g., Knefelkamp & Widick, 1975;

Stephenson & Hunt, 1977; Widick & Simpson, 1978). Moreover, are student affairs practitioners or is student affairs as a profession sophisticated enough to provide interventions designed for one specific position of the Perry Scheme?

Perry (1981) suggested that while precise assessments may be required by experimenters, "rough-hewn groupings of students evidencing dualistic thinking, multiplistic thinking, and relativistic thinking will provide ample base for such differential instruction as is economically possible in most classes" (p. 102). This sentiment is probably equally applicable to other situations faced by practitioners. Moreover, Perry (1981) argued that there is "a problem inherent in making finer measurements with such an ordinal scale" (p. 102) because the scheme says nothing about the distance between positions. Many have used averages to create what appear to be interval scores in order to compare groups and calculate correlations and other statistics that require interval scales. Perry (1981) admitted "we ourselves are among those who have committed this sin against parametric statistics" (p. 102).

Item Response Theory

Although the use of a latent trait model may at first appear attractive for the measurement of development, several problems exist which would make it difficult to apply. Basic item response theory (IRT) assumes that items are scored dichotomously (Weiss & Yoes, 1990), usually based on whether the respondent was correct or incorrect. Forcing a respondent into a yes-or-no, true-or-false situation has implications for item development. Given the complexity of the construct, some might argue that such a dichotomy would restrict the frame of reference of the respondent too much. Although IRT methods exist for multichotomous items, they are much more complex (Weiss & Yoes, 1990).

Although Perry (1970) described the scheme as a unidimensional construct, most describe the scheme as comprised of four underlying dimensions (i.e., dualism, multiplicity, relativism, and commitment). Beyond that, the argument has been made by many that two separate constructs comprise the scheme (described earlier). However, unidimensionality is a critical assumption of item response theory (Allen & Yen, 1979; Crocker & Algina, 1986). Two particular problems arise from this assumption. First, an objective instrument would be required to force a unidimensional structure on the scheme; so far, researchers have not been able to do this. Second, if three or four dimensions of the construct are maintained, the concern becomes within-stage development. As described above, the research has not yet identified specific behaviors associated with within-stages levels of development. Consequently, although empirical item analyses may distinguish between items that represent higher and lower within-stage development, these differentiations may not be supported theoretically.

Finally, the complex nature of the stages must be addressed. IRT techniques identify the probability that an examinee with a certain trait value (e.g., intellectual developmental level) will get a particular item correct, or, more generally, answer it in a particular way (Allen & Yen, 1979). Indeed, "at the 'heart' of the theory is a mathematical model of how examinees at different ability levels for the trait should respond to an item" (Crocker & Algina, 1986). However, the complex stage assumption of the Perry Scheme suggests that this level may differ dependent upon the particular context of the measurement. Unfortunately, many measurement techniques make the same simple stage assumption as IRT by providing a single stage score for individuals. No consensus has yet been reached as to the best scoring methods to use for the Perry Scheme.

Other Measurement Issues

Beyond the psychometric properties of reliability and validity, King and Kitchener (1994) have identified several criteria that can be adapted to the evaluation of instruments for the Perry Scheme: (a) real problems or issues should be the focus of the assessment task; (b) the assessment should elicit information about the rationale for a response as well as its content; (c) thinking should be sampled across a variety of issues; (d) the content should be familiar to a wide range of individuals; (e) the content should not constrain the instrument's use to individuals in formal educational settings; (f) the reading level should be such that it can be used with a wide range of potential test takers; and (g) the theoretical model on which the instrument is based must have been validated.

Sample Selection Issues

Certainly, the best choice of research design to determine the effect of one variable (e.g., college attendance) on another (e.g., cognitive developmental outcomes) is through the random selection and then random assignment of individual subjects to experimental and control groups. Unfortunately, the conditions necessary for a true randomized experiment are nearly impossible to obtain for collegiate studies (Pascarella & Terenzini, 1991). The largest impediment is that individuals choose for themselves whether to attend college, where to attend, and even whether to continue each year. That is, a true outcomes experiment would require randomly assigning individuals to attend or not to attend college. Consequently, other sample selection methods and other research designs are necessary for such inquiry.

Difficulties in Sample Selection

Because the obvious baseline for assessing the effect of college is to compare college graduates to individuals who did not attend college, the straightforward approach would be to gather data from such groups and compare them (Pascarella, 1989). Similarly, longitudinal studies of subjects at the same age who do and do not go to college are helpful in sorting out the effects of age and education (King & Kitchener, 1985). Unfortunately, it is difficult to collect comparable data for individuals whose education stopped after secondary school, because there are few if any naturally occurring and easily accessible groups of noncollege individuals; therefore, an appropriate control group for college outcomes assessment may be very difficult to assemble (Pascarella & Terenzini, 1991).

Even if such a sample could be obtained, these two groups might be so fundamentally different in important characteristics that their comparison is meaningless (Pascarella & Terenzini, 1991). Although the groups can be matched or equated on a small number of variables, or the effects of these variables can be controlled statistically (e.g., partial correlations, analysis of covariance), the basic differences between the groups may still make analysis difficult. .For example, among the group of noncollege persons, some may have had job training and some may have been unemployed. Also, it is difficult to control for reasons of attendance or nonattendance, such as personality or family values about college.

Pascarella and Terenzini (1991) have reported that overwhelming evidence indicates that student characteristics are not randomly distributed among (a) college and noncollege groups, (b) different types of colleges, or even (c) different academic and social experiences within the same institution. This evidence poses the challenge of separating the influence of college from the influences due to other personal characteristics of the students. Because individual student background characteristics influence both the collegiate experience and outcome measures, they satisfy the classical definition of a confounding variable (Pascarella & Terenzini, 1991).

Suggestions for Sample Selection

Although obtaining a sample of individuals who do not attend college may be difficult, there are subpopulations within the campus environment that can provide some insight into specific programmatic areas. For example, Terenzini, Pascarella, and Blimling (1996) have found that students' out-of-class or cocurricular experiences (for example, residence hall programs, organizational involvement, decision making, social activities, internships, employment, structured activities, and interaction with faculty and peers) do influence students in different ways. Indeed, Baxter Magolda (1992a) concluded that epistemological development was a more important factor than class rank for determining how students interpreted their experiences. Also, Rest (as cited in Kitchener, 1982) found that students who live away from home during college show larger increases in moral judgment.

For future research, then, intellectual development of students who live in college residence halls for much of their college career can be compared to students who have lived near campus, have commuted to campus, or have lived at home (i.e., with parents) during college. Differences among student involvement in campus life can be assessed by comparing levels of organizational involvement, for example members of Greek organizations versus those who did not join a Greek house. Students, and especially leaders, who are actively involved in campus organizations with a less social mission (e.g., student government) can be compared to students who are not involved or are less involved. Further, students can be compared based on such personal variables family history, socioeconomic status (e.g., students who must work), enrollment status (e.g., part-time), or support network (e.g., out-of-state).

Also, it is necessary to ensure that students with diverse personal characteristics are present in sufficient number to allow for analyses of possible interaction effects. Some of these characteristics that are potential moderator variables include sex, race, age, socioeconomic background, academic major, academic achievement, personality type, and sexual orientation (Widick, 1977; Winston & Moore, 1991). For example, Kitchener and King (1981) found a weak but significant correlation between socioeconomic status and reflective judgment. Also, Durham, Hays, and Martinez (1994) found that class rank was not a significant contributor to Perry position among first generation Chicano students as it has been for Anglo American students. They concluded that because one's knowledge and perspectives are closely associated with one's social class, the different sociocultural support networks may have been the primary reason for this difference.

On a larger scale, collegiate impact should be studied through comparisons of differing types of academic institutions (Astin, 1978; Erwin, 1983; Widick, 1977). Comparison variables could include such institutional characteristics as geographical region, selectivity of admission, residential nature of the campus, public or private support, and range of degrees offered. For example, graduates of two-year institutions can be compared to seniors or graduates of four-year institutions.

Other populations also may provide valuable information concerning intellectual development, particularly for the refinement of theory. For example, because the college years only capture a small portion of adult development, an examination of the continued development of graduates could be beneficial. College may have an impact even if no apparent development occurs. Pascarella (1991) has argued that typical designs focus only on direct and unmediated effects, ignoring such long-term, indirect effects as occupational positioning (high school graduates do not receive the intellectual stimulation offered by college and also tend to occupy less intellectually stimulating jobs). A comparison of the continued development of graduates with similar-aged individuals who did not attend college might indicate whether college encourages cognitive development even after college. That is, learning how to learn in college may assist further development beyond the college years. Comparisons of this nature also could be made between graduate students and those whose education stopped either after college or after high school.

For example, Rest (as cited in Light, Singer, & Willett, 1990) compared three groups with differing amounts of college eventually completed (college graduates, 2-4 years of college, and under two years) over a 10 year period. Rest found that the three groups tended to follow the same developmental trajectory during the first few years; but after ten years, the high and medium groups had retained their previous moral developmental status, while the low group regressed nearly to their previous high school levels. Certainly there is also much to be learned about intellectual development from the years following college.

Research Design Issues

Traditionally, cross-sectional designs and longitudinal designs have been used in developmental research and also in outcomes assessment (Winston & Moore, 1991). In cross-sectional research, data are collected at the same time from two groups, which are called cohorts. The typical cross-sectional design used in student developmental research compares freshman, who have had no exposure to college, with seniors, who have usually had four years of college impact. Differences between the groups are thought to be a result of the intervention or treatment, which in the case of cognitive developmental research is the influence of college instruction and cocurricular programming.

Longitudinal studies, sometimes called panel studies, typically are designed to observe changes in a single group of individuals over a meaningful period of time. Measurements are taken at the beginning and the end of the time period, providing a comparison of the individuals to themselves at a later point in time. Sometimes repeated measurements are taken at intervals within the time period, often with the hope of showing trends in the developmental process. Sometimes, also, a control group is measured in the same manner as the treatment group; however, it is very difficult to obtain such a control for college student developmental studies, as has been documented above. Both cross-sectional and longitudinal types of designs, however, have shortcomings. Some of the more important considerations for each design choice will be discussed.

Difficulties in Research Design

Many issues can influence developmental research. Some of these confounding factors include practice effects in longitudinal studies, which can occur as a result of facing the same instrument more than once. An individual may be influenced by memories and the opportunity for practice afforded by previous testing situations. Another issue brought about by repeated measurements is that of regression to the mean: through no effect of the intervention, exceptional scores tend to move closer to the mean at subsequent measurements (Baltes, Reese, & Nesselroade, 1988; Hopkins, Stanley, & Hopkins, 1990).

Among the most significant factors that can influence both longitudinal and cross-sectional research are age or maturation, external historical events, personal traits, and attrition.

Age of Students. Much of the development that occurs coincidentally with college attendance may be explained by the fact that students grow older as they attend college; this implies that average freshman-to-senior changes may well overestimate the net influence of the college environment on development (Pascarella & Terenzini, 1991). This is a factor that affects both cross-sectional and longitudinal designs. That is, regardless of the effect of college, individuals who are 21 years of age could be expected to show greater intellectual complexity than 18 year-olds. Likewise, a person measured during freshman year can be expected to appear more complex as a result of growing older, and not necessarily because of college attendance, when measured four years later. Educational level can be a similarly confounding factor, because it is so highly correlated with age.

Attrition. Also a matter of concern for both cross-sectional and longitudinal designs, attrition manifests itself in several ways. For cross-sectional studies, comparisons between freshmen and seniors may be inadequate because not all freshmen continue in college to their senior years. This type of attrition may cause the seniors to be a more select group on key variables related to the developmental process, thereby making comparisons biased.

Two types of attrition may affect longitudinal studies. First, students who are measured in their freshmen year, for example, may not continue in college until the second measurement of the sample is taken. Also, some who do persist through college may choose not to participate in the follow-up portion of the research. As a result, what may have begun as a representative group no longer is, thereby diminishing the generalizability of the study. Moreover, because the follow-up group is rarely comprised of all the persons who were measured originally, longitudinal studies often require very large initial samples. Second, particularly in studies where the measurement techniques require subjective analysis, not all researchers may be able to continue throughout the entire research project.

History and Personal Traits. Beyond the issue of age, several other factors make it difficult to ensure that the groups are equivalent to each other in all ways except the treatment intervention (Winston & Moore, 1991). For example, Pascarella and Terenzini (1991) noted that because colleges tend to recruit and enroll different types of students and because students are more or less involved in campus activities, it is difficult to separate the effects of college from the influence of individual differences. Other personal characteristics that may confound the results include personal ambition, academic preparation, learning style, attendance patterns, motivation to succeed, career aspirations, and other personality variables (Pascarella, 1991).

Historically, recruitment and admission criteria may change between classes, as well as other historical events that may cause the groups to differ in some substantial way. Cultural and social changes or significant historical events may somehow influence the way students approach learning and college. For example, Schmidt (1985) measured a group of traditional freshmen beginning one year and did not measure the nontraditional group until the following year. Therefore, in order to perform a "cohort check" for the first group, she measured another traditional freshman group along with the nontraditional group to determine if the simple passage of time had caused any differences.

Other Issues Concerning the Basic Designs. Most longitudinal studies do not make use of an independent control group. Therefore, although single group longitudinal studies are able to document changes and growth, one cannot be certain what caused the changes (Spector, 1981). Also, there are rarely baselines with which to compare whether the development was beyond an expected amount. Additionally, some of the assessment problems described above, especially in regard to the slowness of development and within-stage development, make traditional time-series longitudinal designs and case studies difficult. These designs are better suited for the analysis of the effects of particular treatments or interventions. Change in developmental status, which is the result of a slow process and not individual interventions, often remains the same after such programming efforts.

The issue of process versus status described above also influences decisions regarding research design. Cross-sectional designs are potentially useful in showing the developmental differences expected between groups. However, in the study of development specifically, cross-sectional studies are flawed fundamentally because they compare groups, one of which has not yet developed and one that allegedly has. More specifically, they do not measure the change within individuals, where development actually occurs. Light, Singer, and Willett (1990) have noted that many researchers try to examine development by using data from only two points in time; however, they argue that the nature of development requires that measurements be taken at more than two times. Among the benefits to measuring development on at least three occasions is the ability to characterize the trajectory of growth (e.g., linear, exponential, curvilinear).

Suggestions for Research Design

Baltes, Reese, and Nesselroade (1988) have indicated that the three major conventional methods of data collection for developmental research are: cross-sectional, longitudinal, and time-lag. Cross-sectional and longitudinal designs have been described briefly above. Brabeck (1984) summarized the time lag design as a sequential study in which independent samples are drawn at different times such that the second group is comprised of individuals who are the age that the first group's members would have been if they were participating in a longitudinal study.

Because the problems described above, especially the age and history factors, affect all three of these primary research designs, scholars have suggested additional methods that better account for both interindividual and intraindividual change. Baltes, Reese, and Nesselroade (1988) have provided detailed descriptions of two of the best designs available for developmental research. These designs have been called sequential strategies. Although primarily used in long-term (e.g., several decades) research projects, they can be adapted to the shorter term required for studies of college student development.

The first sequential design is called the cross-sectional sequence (see Figure 1). Briefly, in the cross-sectional sequence design, independent observations are obtained for all cohort (i.e., year of birth) and age (e.g., class rank) levels. For example, Pascarella and Terenzini (1991) recommended a cross-sectional design based on four groups of different age-rank characteristics: traditional 18 year old freshmen, traditional 21 year old seniors, nontraditional freshmen (perhaps age 21), and nontraditional seniors (e.g., age 25). In such a design, the effects of age could be obtained by comparing average developmental levels of students of the different ages but with the same amount of exposure to college (i.e., class rank); the influence of college independent of age would be obtained by comparing the groups with different class ranks but the same age. Even with such a cross-sectional design that controls for the age of students so well, changes in developmental levels are not guaranteed to be caused by college attendance since other factors may still confound the results. In particular, history effects, or sometimes called cohort effects, can influence the study. Also, Baltes et al. noted that cross-sectional sequences can provide only average intraindividual change data.

The second sequential design is called a longitudinal sequence (see Figure 1). Essentially, two cohorts are identified and repeated measures are obtained within each cohort. For example, a freshman class would be sampled and then follow-up measurements taken during senior year, or possibly also at interim intervals. Then, for another freshman class the procedure would be repeated. This longitudinal-sequential method that measures two or more groups of cohorts simultaneously and in repeated measurements permits inferences about age change and cohort differences and is "ideal for the task of direct and precise descriptions of intraindividual and interindividual components of developmental change" (Brabeck, 1984, p. 16). Clinchy and Zimmerman (1982) used such an approach in their studies of undergraduate women because the design allows longitudinal, cross-sectional, and time-lag comparisons that allow researchers to separate out effects of age, cohort, and time of measurement (i.e., historical or cultural changes).

Qualitative Methodologies. Other variables, particularly environmental variables, must be examined for their impact on student development. For example, teaching styles, teaching effectiveness, classroom environments, and residence halls and other campus facilities each probably influence student development in some way. However, current methods for assessing student development tend to ignore these factors. Although some environmental assessment instruments exist, they are not designed from a Perry Scheme and outcomes assessment perspective. Therefore, perhaps the best way to begin to study these factors is from a qualitative approach.

Specifically, qualitative methods share a primarily phenomenological perspective that emphasizes understanding the meaning that events have for the persons being studied (Patton, 1991). Patton indicated that quantitative approaches are based on logical positivism, the perspective of the natural sciences that emphasizes the acquisition of objective knowledge. Patton (1991) provided an outline of four critical distinctions between the two methodological approaches. First, the quantitative and qualitative approaches differ in their assumptions of ontology, or what constitutes the social world. Second, the two approaches differ about how individuals know what they know about the world, called epistemology. Third, the two approaches differ in regard to their goals and objectives of inquiry into the behavior of persons. Finally, the two approaches differ in the methods they use for data collection and analysis.

Using both quantitative and qualitative methods in studying a single phenomenon is commonly called method triangulation (Caple, 1991; Stage & Russell, 1992). Some argue that to use both types of methods is risky because they are designed to answer different questions and most researchers are not equally proficient with both (Caple, 1991). However, different methodologies offer different strengths and weaknesses and the biases or limitations inherent in any particular method may be compensated for by the counter-balancing strengths of another method, resulting in a convergence toward reality (Caple, 1991; Stage & Russell, 1992).

Stage and Russell (1992) have argued that the strategy of implementing multiple methods provides greater accuracy and completeness in describing a phenomenon. By viewing an issue from a variety of perspectives, more sides of it are exposed. Triangulation views qualitative and quantitative methods as complementary and that the methods used together may improve the validity of the results. Further, triangulation can enrich conclusions by providing evidence for elements of a phenomenon which may go unnoticed using a more standard approach. Unfortunately, in addition to convergence of results, triangulation can also produce inconsistent or contradictory results (Patton, 1991; Stage & Russell, 1992).

The primary concern of the research process in student affairs is the notion of how we learn about our students, particularly lesser studied subgroups of student like women and minorities (Hanson, 1990). The quantitative approach may provide data about many people, thereby facilitating comparisons between groups of people on a limited number of questions. Qualitative methods usually produce a wealth of detailed information about a much smaller number of cases, for which the analyses are much less generalizable. However, the detailed and more sensitive information can help to illuminate subtle complexities of the collegiate experience that would inaccessible through quantitative methods (Pascarella, 1991). For example, observations of students in class or in social contexts could help researchers identify subtle differences men and women experience in the learning environment and also types of involvement that aid or hinder intellectual development (Baxter Magolda, 1990). Similarly, qualitative observation could help answer the question of what causes developmental movement.

Recommended Research Agenda

Based on (a) the recommendations of those who study development, (b) on the difficulties of planning developmental research, and (c) on the problems that have beset previous efforts in studying the Perry Scheme in particular, the following research agenda is recommended. Several research priorities will be described through which it can be determined whether the Perry Scheme has a valid role in outcomes assessment. First, a study will define the parameters by which an objective instrument can be developed to measure the Perry Scheme. Second, studies will be presented through which several remaining pieces of critical validation evidence can be provided. Finally, studies will be described by which the assessment of outcomes, or value-added, can be established.

Research Priority 1: Instrumentation

Several possible measures exist for assessment of development according to the Perry Scheme. Unfortunately, the most well-established techniques are subjective methods. The point was made earlier that much of the validation work still required for the Perry Scheme involved large-scale assessment across diverse populations of students and institutions, which precludes all the subjective methods as possibilities. The subjective methods, including the MID and the MER, require too much time for both administration and rating. Further, the subjectivity of the rating process would make results across large-scale administrations difficult to verify.

Therefore, an objective instrument must be identified. Unfortunately, no objective method has been established as reliable and valid through multiple studies. The SID was validated initially using an inappropriate sample (too many freshmen) and subsequent testing by other researchers found that only the SID-Dualism scale showed possible validity as a measure of one dimension of the Perry Scheme. No evidence has been reported for the content validity of the SID, and construct validity was inadequate--one factor (Empathy) was defined that bears no explained relationship to underlying concepts of the Perry Scheme. The SID may yet become a valuable Perry assessment if the empathy factor is removed and further testing is then done on the remaining factors. Also, more specific position ratings will make the SID more useful.

Even where the psychometric properties are known, the PCDI, LCQ, Foster's Checklist, and Ryan's instrument have not been used by enough researchers to determine their usefulness. The SAID and the WOKI, although they show promise, are not measures of the Perry Scheme per se. Similarly, Fago's scale shows some potential but must be used and tested more extensively.

Of all the objective instruments, Moore (1989) has provided for the LEP the most detailed description of instrument development and the most adequate evidence of validity. Therefore, if an instrument was needed for immediate use, the LEP would be recommended. However, the initial validation sample was also too heavily weighted with underclassmen. Factor analysis provided less than desirable evidence of construct validity, with several factors not well-defined. Additionally, the LEP is limited to positions 2-5, which may limit its usefulness for outcomes assessment. Criterion validity was assessed through comparisons to the MID, but the correlations were lower than what might have been expected, especially considering that the LEP was developed based on the MID.

Development of a New Objective Instrument

The assessment and measurement problems cited earlier for the development of an instrument to measure cognitive development based on the Perry Scheme are numerous. For example, many opinions exist regarding the issues of complex versus simple stages and appropriate scoring procedures. No single measurement technique can address them to the satisfaction of all who would use a measure of the Perry Scheme. This section, therefore, will address only the most important issues for the development of an objective measure of the Perry Scheme.

If an objective instrument can be developed that maintains some of the nature of a production task, it certainly would appease many in the Perry community who believe that to be the sine qua non for the assessment of the Perry Scheme. For example, King and Kitchener (1994) are making an effort to create a more easily administered and scored instrument, which they call the Reflective Thinking Appraisal (RTA). In this approach, respondents are presented an ill-structured problem and asked to respond to a series of questions similar to standard probe questions of the RJI. The critical feature is that respondents will be asked to briefly write down their own answer to each question, which elicits a production of response from the respondents. Then, respondents are asked to read several prototypic responses to the question and rate each on its similarity to their own; these rankings are used to determine an individual's score. Unfortunately, this instrument has apparently been very difficult to develop, given that King (1990) noted that development of the instrument had begun in 1983. For example, the production aspect of the instrument could be easily faked by paraphrasing the given prototypic responses.

Therefore, an approach similar to Moore's (1989) development of the LEP seems appropriate. Also, the validation evidence, especially from the SID, the LEP, and the SAID, shows that existing Dualism scales may be strong. It may be possible to adapt items from these instruments to fulfill the objectives of the new instrument. The MER has been the subject of intense, and more recent, investigation by Baxter Magolda and her colleagues. The development of the rating manual has proceeded in such a way so as to standardize the rating system as much as possible. This standardization should assist in the development of an objective instrument. Items could be constructed from each domain of the MER and for each position based on the rating manual. Although some evidence has suggested that Position 1 may not be common among college students, it should be included in the instrument for possible use with high school students or pre-college orientation sessions. Similarly, wording of the items should be general enough to be applicable to noncollege individuals.

Further, the theoretical refinements that have been possible as a result of the MER research are important to consider in a new instrument. In particular, attention must be paid to the parallel development of the relational and impersonal patterns of reasoning. Also, based on earlier arguments, at least one position of Commitment (i.e., Position 6 or Position 7) should be included, specifically addressing commitments to opinions and the creation of knowledge. Inclusion of Commitment positions will also make the instrument more amenable to use with graduate students and other older persons not studied much in previous Perry Scheme research. Meaningless items could also be added to account for a possible social desirability response set. Since Moore's procedure has been described earlier, it will not be repeated here. The resulting Likert-type instrument could be called the Hypothetical Objective Measure of Epistemological Reflection (HOMER).

The HOMER should be tested for content validity by having several experts (unlike Erwin, 1983, who apparently performed this task alone), in both the Perry Scheme and the MER, review the items for relevance to the constructs and domains. The first evidence of construct validity for the HOMER should be established through factor analysis. Given the interaction of domains and positions, the factor structure may be somewhat convoluted. However, the factors should reasonably represent Positions 1 through 6/7 of the Perry Scheme. After the factor analysis is performed and evaluated, along with the reliability analyses and item analyses, poor items should be removed or reworded. A second validity study should be performed based on the redesigned instrument, including a second factor analysis. Only the development of the SAID has described such a procedure. At that point, if strong factors can be verified, further validation studies should be attempted. Each factor should be scored individually, like the SID. If growth occurs in longitudinal studies, the lower position factors should decrease in score while the higher positions increase. The factor with the highest score would be considered the dominant position. This will help maintain a complex stage model approach to scoring.

Criterion validity should be assessed through comparisons with the MER or the MID for every validation study participant. In addition, a random sample of participants should be interviewed to provide additional support. The MID and the MER have shown reasonable correlation to interviews, but most have agreed that interviews are the best means by which to assess Perry position. One other reason is that the MER and the MID only measure as high as Position 5, but items about commitment to opinions and the creation of knowledge would be included on the HOMER for Commitment Position 6/7. Further, it may be possible through such comparisons to determine whether recognition formats do indeed produce higher scores than production task formats.

Additional construct validity should be established through convergent and discriminant validity. In particular, a scale of moral development, psychosocial development, identity development, or career development should be related to intellectual development because several researchers have found such relationships to exist. Discriminant validity can be assessed through a number of alternatives, such as critical thinking, intelligence, or personality. Another test of construct validity is to examine the expected increases based on class rank through an analysis of variance. That is, freshmen should score lower than seniors. Further validation of the instrument will occur as the theory itself is validated through studies described later.

An Alternative Approach

The procedure described above may fall victim to the same apparent lack of success that has befallen past attempts to measure the Perry Scheme using a Likert-type scale. It may be worthwhile to attempt another scale format, such as an adaptation of Thurstone scales of equal appearing intervals. In particular, a perspective change from a production view to an attitude view may provide fertile ground for future research. That is, cognitive development may be described as an attitude toward the nature of knowledge. Indeed, many items and rating criteria of instruments and measurement processes described earlier identify such attitudes as a basis for rating a student's development status. Perhaps shifting perspective to an attitude-toward-knowledge point of view can help overcome some of the problems that have been faced thus far in the development of quantitative Perry instruments.

Items can be created in the same way using the MER as a foundation as described above. However, the experts must act as judges to rate the items. Thurstone apparently required many judges to rate items, but Babbie (1989) and Mueller (1986) have suggested that 10 to 15 judges should suffice. Although not a true Thurstone approach (which measures intensity of attitudes), the judges would rate the items from 1 to 6 for the Perry position they represent. The items would each receive a scale value based on the median of the judges' ratings. Those items which do not receive a near consensus would be removed. Items would be selected based on their scale values and incorporated into a scale, perhaps called the Thurstone Attitude Scale about Knowledge (TASK).

Respondents would select and mark the items on the TASK with which they agree and their score would be the average scale values for the items they choose. Additional information would be obtained by the highest and lowest scale values they chose. Of particular interest may be the lowest scale value item the respondent chooses; it may indicate what level the respondent is certainly above, whatever the average level may be. An increase over time of this lowest level may be as interesting and informative as an increase in the average level. Meaningless items could also be added to the TASK to account for the potential of a social desirability response set. Also, validity and reliability studies could proceed in the same manner as described above.

Although the Thurstone technique is not an IRT method, it does provide "item scores" that make interpretation of the instrument a very straightforward process. The Thurstone technique is particularly time-consuming and labor-intensive during development, particularly in relation to the judging process. However, there are many academicians and practitioners familiar with the Perry Scheme who may be willing to assist with such a project.

Research Priority 2: Validation of the Perry Scheme

Study 1: Diversity

The second research priority is to provide further evidence of the validity of the Perry Scheme. This validity evidence must be obtained in respect to the specific issue of diverse populations of students and institutions. Therefore, the purpose of Study 1 will be to gather data on a large scale from a variety of institutions and a diversity of individuals. Two design issues must be considered for such a study: (a) both cross-sectional and longitudinal evidence is required to show that a developmental theory is valid and (b) the large scale of the study requires practical methods. It would be very difficult to obtain longitudinal data on such a large scale. However, the study can be broken into two phases. First, large-scale cross-sectional data would be obtained. Measurements must be taken across class rank, such that freshmen, sophomores, juniors, seniors, and graduate students are included from each institution that has the appropriate academic ranks. Second, within a subset of the original institutions, repeated longitudinal data should be gathered where possible; otherwise, longitudinal data with independent observations would be collected (that is, data would be collected over four or more years from a new sample of students each year). Given the scope of this study, the cost of sequential designs would be prohibitive.

Institutions must be identified that will provide a broad range of institutional characteristics and student characteristics. More specifically, (a) institutions from more than one geographic region of the country should be chosen, including rural and urban, commuter and residential campuses, (b) insitutions chosen should represent comprehensive universities, community or two-year colleges, liberal arts colleges, historically predominantly black institutions, and probably an institution from a southwestern state with a relatively large Hispanic and/or Asian student population, and (c) both public and private institutions should be included. Random sampling within such an array of institutions should provide an adequate mix of student characteristics. However, some form of stratified sampling would ensure that particular student variables would be included, most notably race, age, sex, and socioeconomic status. After the data are collected, analyses should be performed to make the appropriate determinations regarding the validity of the Perry Scheme for diverse students and across a variety of institutions.

Study 2: College Impact

Noncollege Control Group. The next piece of validation evidence required is the comparison of college students to persons who finished their formal education with their graduation from high school. Inasmuch as Perry described his theory for college students, it must be shown that college students differ in intellectual development from noncollege individuals. Therefore, a sample must be obtained from among the population of persons who have not attended college. Although similar considerations as described above would be beneficial (e.g., geographic variety, occupational variety), a diversity in the primary demographic variables of age, race, sex, and socioeconomic status must be well represented. A multiple-group multiple-measure design would be used to compare college students to noncollege persons over, for example, a ten-year period, with measurements taken every two years. Although such a study would be exceedingly difficult, it is the only true means by which the Perry Scheme can be validated. Pascarella (1991) argued that unless the changes that occur in people who do not attend college are known, researchers may be misled by change they observe among college students.

Specifically, if the Perry Scheme cannot illuminate developmental differences between college and noncollege individuals, it must be considered inappropriate to use for outcomes assessment.

If such a longitudinal design cannot be achieved, a cross-sectional design may have to suffice. For example, last-semester college seniors can be compared to 22-year old individuals who did not attend college. Although such a sampling strategy may prove more practical, it leaves problem areas for the analysis, such as what were the developmental levels of the respective groups when they graduated from high school. A third group, namely students who attended college only 2 or 3 years (e.g., withdrawals or 2-year college students), may help show a developmental progression based on the number of years in college. A comparison with high school seniors or orientation freshmen may provide baseline information, especially if the high school seniors can be grouped based on whether they plan to attend college.

Sequential Design. A valiant attempt must be made to obtain a sample of persons whose education stopped after post-secondary school. However, if such a sample proves too difficult to obtain, other designs will be required. The sequential designs described earlier are the best choice for such a purpose. For example, a longitudinal sequential design could be used to compare students of different ages. Or traditional and nontraditional students could comprise two cohorts of students who are measured simultaneously over the course of their college careers, perhaps even through graduate school. Such a design would help account for the usually confounding factor of age. Once again, if such a longitudinal study proves too difficult, a longitudinal sequence with independent observations or a cross-sectional sequence design can be used. However, one of these designs must be used to account for the maturation process.

Study 3: Causes of Development

Two other pieces of validity evidence that have been suggested in the literature are (a) more specificity of the causes of developmental progress and (b) more specificity of the behaviors associated with particular developmental levels. Unfortunately, typical objective measurement techniques do not access these areas. Indeed, Kitchener (1985) reported that asking students to recount such information retrospectively in interviews has not worked either. Therefore, qualitative methods should be employed to better understand what it is about college that causes student development. Then, through analysis of the data, these factors must be related to the Perry Scheme.

Several possible methods may be useful for these purposes. For example, a sample of students may be enlisted to maintain journals (for credit of course) throughout their freshman years. These journals could be analyzed after the year for catalysts that caused developmental movement, or possibly deflections from growth. Such a study may provide information regarding different levels of within-stage development. Further, particular behaviors associated with the developmental levels may be identified. This study should be done in conjunction with at least beginning and end of year assessments (preferably at least three measurements) of development using the HOMER or the TASK.

Another qualitative study that could be performed is a weekly observation of classroom experiences over a quarter or the year. Such a study could identify teaching behaviors and other environmental factors that influence development. Further, it would be possible to identify behaviors that are utilized by individuals at specific Perry positions. Of course, this would require multiple assessment of Perry position using the TASK or the HOMER as well. Further, possible within-stage development may be identified along with catalysts for growth or deflection. Similar studies could be performed through the regular observation of other critical areas of campus life, such as residence life or student organizations.

Research Priority 3: The Perry Scheme for Outcomes Assessment

The preceding research priorities will help establish the Perry Scheme as a valid theory for use with outcomes assessment. Also, validity evidence will be provided for either the HOMER or the TASK. However, pilot studies must be performed to determine whether the Perry Scheme and the HOMER or the TASK will be practical as tools of outcomes assessment. Therefore, several recommendations will be made to test outcomes assessment applications. These particular recommendations have been made also to provide additional validity evidence for the Perry Scheme.

Based on the levels of involvement that would be associated with different residential statuses, one study could examine the differences among residential students, near-campus residents, off-campus commuter students, and live-at-home students. Specifically, freshmen of each group should be measured upon entry to college (perhaps during orientation) and then again at the end of the freshman year. Those who maintained the same residential status during sophomore year could be recruited to continue with the study through the end of sophomore year. It would be unlikely that many students would continue with the same living arrangements throughout their four years, and therefore sophomore year may be the practical limit of this study.

Also, several researchers have determined that students who are actively involved in student activities and organizations show more development during college. This finding could be tested relative to development according to the Perry Scheme by comparing groups of student leaders over a given year. Many student organizations hold elections during the prior spring. These students could be identified and measured either then or at the beginning of the fall. A group of students who are not involved in student organizations could be recruited to participate in addition to less-active members of some of the organizations. Again, measurements would be required at the end of the year. This study could provide some insight into the quality or quantity of student involvement required to facilitate development. However, some issues (like age and class rank) will confound the results if not controlled either through the sampling method or through statistical methods.

Another study could examine the relationship of support networks to cognitive development. Some evidence has shown that support networks are important to the developmental process. This could be examined from a Perry perspective by comparing a group of in-state students to out-of-state students. Another study could investigate the relationship of peer support networks through comparisons of students who belong to social organizations (e.g., fraternities and sororities, clubs) to those who do not.

One final area that could produce promising information is the comparison of different levels of interaction with university faculty and staff (e.g., student personnel professionals, organization advisers). Students could be identified by faculty and staff for levels of interaction and measured with the HOMER for some period of time. Alternatively, students can be selected based on self-reports of their involvement with faculty or staff.

Final Notes

Until we have reliable normative data about intellectual development (if we ever do), outcomes assessment must continue to examine longitudinal data of some kind. Although validity studies require the use of repeated longitudinal measurements on the same group of students, the needs of outcomes assessment may allow less restrictive procedures. For example, entering freshmen, either the entire class or a sample, may be requested to respond to a Perry Scheme measure, such as the HOMER or the TASK, as part of the standard orientation process. Then seniors may be recruited to complete the instrument before they graduate. A similar procedure is used by many universities with the ACT-COMP. The class averages can then be used (or in some cases actual repeated measures) to determine the value added by college attendance, at that particular college, in the area of intellectual development.

As Hanson (1990) argued, we must assess the manner in which the process of higher education influences students to learn and grow; specifically, we must determine what aspects of the environment and activities facilitate development and thereby link how students develop to the programs and services offered on a college campus. In short, we need to determine specifically what it is about college that helps students develop and what it is to which persons who do not attend college are not exposed.

REFERENCES AND BIBLIOGRAPHY

Adams, G. R., Shea, J., & Fitch, S. A. (1979). Toward the development of an objective assessment of ego-identity status. Journal of Youth and Adolescence, 8, 223-237.

Alishio, K. C., & Schilling, K. M. (1984). Sex differences in intellectual and ego development in late adolescence. Journal of Youth and Adolescence, 13, 213-224.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole.

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan.

Astin, A. W. (1978). Four critical years. San Francisco: Jossey-Bass.

Astin, A. W. (1987). Assessment, value-added, and educational excellence. In D. F. Halpern (Ed.), New directions for higher education: No. 59. Student outcomes assessment: What institutions stand to gain (pp. 89-107). San Francisco: Jossey-Bass.

Babbie, E. (1989). The practice of social research (5th ed.). Belmont, CA: Wadsworth.

Battaglini, D. J., & Schenkat, R. J. (1987). Fostering cognitive development in college students--the Perry and Toulmin models. Urbana, IL: ERIC Clearinghouse on Reading and Communication. (ERIC Document Reproduction Service No. ED 284 272)

Baxter Magolda, M. B. (1987). A rater-training program for assessing intellectual development on the Perry Scheme. Journal of College Student Personnel, 28, 356-364.

Baxter Magolda, M. B. (1987). Comparing open-ended interviews and standardized measures of intellectual development. Journal of College Student Personnel, 28, 443-448.

Baxter Magolda, M. B. (1988). Measuring gender differences in intellectual development: A comparison of assessment methods. Journal of College Student Development, 29, 528-537.

Baxter Magolda, M. B. (1989). Gender differences in cognitive development: An analysis of complexity and learning styles. Journal of College Student Development, 30, 213-220.

Baxter Magolda, M. B. (1990). The impact of the freshman year on epistemological development: Gender differences. Review of Higher Education, 13, 259-284.

Baxter Magolda, M. B. (1992a). Cocurricular influences on college students' intellectual development. Journal of College Student Development, 33, 203-213.

Baxter Magolda, M. B. (1992b). Knowing and reasoning in college: Gender-related patterns in students' intellectual development. San Francisco: Jossey-Bass.

Baxter Magolda, M. B. (1995). The integration of relational and impersonal knowing in young adults' epistemological development. Journal of College Student Development, 36, 205-216.

Baxter Magolda, M. B., & Porterfield, W. D. (1988). Assessing intellectual development: The link between theory and practice (ACPA Media Publication No. 47). Alexandria, VA: American College Personnel Association. (ERIC Document Reproduction Service No. ED 324 617)

Baxter-Magolda, M. & Porterfield, W. D. (1985). A new approach to assess intellectual development on the Perry Scheme. Journal of College Student Personnel, 26, 343-351.

Belenky, M. F., Clinchy, B. M., Goldberger, N. R., & Tarule, J. M. (1986). Women's ways of knowing: The development of self, voice, and mind. New York: Basic.

Borg, W. R., & Gall, M. D. (1983). Educational research: An introduction (4th ed.). New York: Longman.

Brabeck, M. M. (1984). Longitudinal studies of intellectual development during adulthood: Theoretical and research models. Journal of Research and Development in Education, 17(3), 12-27.

Brown, R. D., & Barr, M. J. (1990). Student development: Yesterday, today, and tomorrow. In L. Moore (Ed.), New directions for student services: No. 51. Evolving theoretical perspectives on students, 51 (pp. 83-92). San Francisco: Jossey-Bass.

Buczynski, P. L. (1993). The development of a paper-and-pencil measure of Belenky, Clinchy, Goldberger and Tarule's (1986) conceptual model of women's ways-of-knowing-instrument. Journal of College Student Development, 34, 197-200.

Cameron, S. W. (1984). The Perry Scheme: A new perspective on adult learners. Syracuse, NY: Syracuse University. (ERIC Document Reproduction Service No. ED 244 698)

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105. (Reprinted in Payne, D. A., & McMorris, R. F. (1967). Educational and psychological measurement: Contributions to theory and practice (pp. 105-114). Waltham, MA: Blaisdell.)

Caple, R. B. (1991). Expanding the horizon. Journal of College Student Development, 32, 387-388.

Clinchy, B., & Zimmerman, C. (1982). Epistemology and agency in the development of undergraduate women. In P. J. Perun (Ed.), The undergraduate woman: Issues in educational equity (pp. 161-181). Lexington, MA: Lexington Books.

Crocker, L., & Algina, J. (1986). Introduction to classical & modern test theory. Fort Worth: Holt, Rinehart and Winston.

Davison, M. L., King, P. M., Kitchener, K. S., and Parker, C. A. (1980). The stage sequence concept in cognitive and social development. Developmental Psychology, 16, 121-131.

Durham, R. L., Hays, J., & Martinez, R. (1994). Socio-cognitive development among Chicano and Anglo American college students. Journal of College Student Development, 35, 178-182.

Ebel, R. L., & Frisbie, D. A. (1986). Essentials of educational measurement (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

ERIC on CD-ROM: 1966-1984 [CD-ROM]. (1997). Baltimore, MD: National Information Services Corporation (Producer and Distributor).

ERIC on CD-ROM: 1985-December 1996 [CD-ROM]. (1997). Baltimore, MD: National Information Services Corporation (Producer and Distributor).

Erwin, T. D. (1983). The Scale of Intellectual Development: Measuring Perry's scheme. Journal of College Student Development, 24, 6-12.

Erwin, T. D. (1991). Assessing student learning and development: A guide to the principles, goals, and methods of determining college outcomes. San Francisco: Jossey-Bass.

Fago, G. C. (1995). A scale of cognitive development: Validating Perry's Scheme. (ERIC Document Reproduction Service No. ED 393 862)

Fleiss, J. L. (1981). Statistical methods for rates and proportions (2nd. ed.). New York: John Wiley & Sons.

Green, M. (1989). Theories of human development. Englewood Cliffs, NJ: Prentice-Hall.

Grotevant, H. D., & Adams, G. R. (1984). Development of an objective measure to assess ego identity in adolescence: Validation and replication. Journal of Youth and Adolescence, 13, 419-438.

Hanson, G. R. (1982). Critical issues in the assessment of student development. In G. R. Hanson (Ed.), New directions for student services: No. 20. Measuring student development (pp. 47-63). San Francisco: Jossey-Bass.

Hanson, G. R. (1990). Improving practice through research, evaluation, and outcomes assessment. In M. J. Barr & M. L. Upcraft (Eds.), New futures for student affairs: Building a vision for professional leadership and practice. San Francisco: Jossey-Bass.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). Educational and psychological measurement and evaluation (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Jones, H. J., Newman, I., Cochran, J. R., & Nemec, W. E. (1992, October). Perry's scheme of intellectual and ethical development: It's implications and correlates in a vocationally undecided population. Paper presented at the meeting of the Mid-Western Educational Research Association, Chicago, IL.

King, P. M. (1978). William Perry's theory of intellectual and ethical development. In L. Knefelkamp, C. Widick, and C. A. Parker (Eds.), New directions for student services: No. 4. Applying new developmental findings (pp. 35-51). San Francisco: Jossey-Bass.

King, P. M. (1990). Assessing development from a cognitive-developmental perspective. In D. G. Creamer (Ed.), College student development: Theory and practice for the 1990s (ACPA Media Publication No. 49, pp. 81-98). Alexandria, VA: American College Personnel Association.

King, P. M., & Baxter Magolda, M. B. (1996). A developmental perspective on learning. Journal of College Student Development, 37, 163-173.

King, P. M., & Kitchener, K. S. (1985). Reflective judgment theory and research: Insights into the process of knowing in the college years. Paper presented at the meeting of the American College Personnel Association, Boston, MA. (ERIC Document Reproduction Service No. ED 263 821)

King, P. M., Kitchener, K. S., & Wood, P. K. (1985). The development of intellect and character: A longitudinal-sequential study of intellectual and moral development in young adults. Moral Education Forum, 10, 1-13.

Kitchener, K. S. (1982). Human development and the college campus: Sequences and tasks. In G. R. Hanson (Ed.), New directions for student services: No. 20. Measuring student development (pp. 17-45). San Francisco: Jossey-Bass.

Kitchener, K. S. (1983). Cognition, metacognition, and epistemic cognition: A three-level model of cognitive processing. Human Development, 26, 222-232.

Kitchener, K. S. (1986). The reflective judgment model: characteristics, evidence, and measurement. In R. A. Mines & K. S. Kitchener (Eds.), Adult cognitive development: Methods and models (pp. 76-91). New York: Praeger.

Kitchener, K. S., & King, P. M. (1981). Reflective judgment: Concepts of justification and their relationship to age and education. Journal of Applied Developmental Psychology, 2, 89-116.

Kitchener, K. S., & King, P. M. (1990). The reflective judgment model: Ten years of research. In M. L. Commons, C. Armon, L. Kohlberg, F. A. Richards, T. A. Grotzer, & J. D. Sinnott (Eds.), Adult Development: Vol. 2. Models and methods in the study of adolescent and adult thought (pp.63-78). New York: Praeger.

Knefelkamp, L. L., & Slepitza, R. (1976). A cognitive-developmental model of career development--an adaptation of the Perry Scheme. Counseling Psychologist, 6(3), 53-58.

Knefelkamp, L. L., Widick, C. C., & Stroad, B. (1976). Cognitive-developmental theory: A guide for counseling women. Counseling Psychologist, 6(2), 15-19.

Kuk, L. (1990). Perspectives on gender differences. In L. Moore (Ed.), New directions for student services: No. 51. Evolving theoretical perspectives on students, 51 (pp. 25-36). San Francisco: Jossey-Bass.

Kurfiss, J. (1977). Sequentiality and structure in a cognitive model of college student development. Developmental Psychology, 13, 565-571.

Kurfiss, J. (1983). Intellectual, psychosocial, and moral development in college: Four major theories. Washington, DC: Council for Independent Colleges. (ERIC Document Reproduction Service No. ED 295 534)

Lenning, O. T. (1980). Assessment and evaluation. In U. Delworth & G. R. Hanson (Eds.), Student services: A handbook for the profession (pp. 232-266). San Francisco: Jossey-Bass.

Lenning, O. T. (1989). Assessment and evaluation. In U. Delworth & G. R. Hanson (Eds.), Student services: A handbook for the profession (2nd ed., pp. 327-352). San Francisco: Jossey-Bass.

Light, R. J., Singer, J. D., & Willett, J. B. (1990). By design: Planning research on higher education. Cambridge, MA: Harvard University.

Martin, J. E., Silva, D. G., Newman, J. H., & Thayer, J. F. (1994). An investigation into the structure of epistemological style. Personal and Individual Differences, 16, 617-629.

Miller, T. K. (1982). Student development assessment: A rationale. In G. R. Hanson (Ed.), New directions for student services: No. 20. Measuring student development (pp. 5-15). San Francisco: Jossey-Bass.

Mines, R. A. (1982). Student development assessment techniques. In G. R. Hanson (Ed.), New directions for student services: No. 20. Measuring student development (pp. 65-91). San Francisco: Jossey-Bass.

Mines, R. A. (1985). Measurement issues in evaluating student development programs. Journal of College Student Personnel, 26, 101-106.

Mines, R. A. (1986). Methodological considerations in young adult cognitive development research. In R. A. Mines & K. S. Kitchener (Eds.), Adult cognitive development: Methods and models (pp. 134-146). New York: Praeger.

Mines, R. A., King, P. M., Hood, A. B., & Wood, P. K. (1990). Stages of intellectual development and associated critical thinking skills in college students. Journal of College Student Development, 31, 538-547.

Moore, L. V., & Upcraft, M. L. (1990). Theory in Student Affairs: Evolving perspectives. In L. Moore (Ed.), New directions for student services: No. 51. Evolving theoretical perspectives on students, 51 (pp. 3-23). San Francisco: Jossey-Bass.

Moore, W. S. (1982). William Perry's cognitive-developmental theory: A review of the model and related research. Unpublished pre-publication draft.

Moore, W. S. (1989). The Learning Environment Preferences: Exploring the construct validity of an objective measure of the Perry Scheme of intellectual development. Journal of College Student Development, 30, 504-514.

Moore, W. S., & Hunter, S. (1993). Beyond "mildly interesting facts": Student self-evaluations and outcomes assessment. In J. MacGregor (Ed.), New directions for teaching and learning: No. 56. Student self-evaluation: Fostering reflective learning (pp. 65-82). San Francisco: Jossey-Bass.

Mueller, D. J. (1986). Measuring social attitudes: A handbook for researchers and practitioners. New York: Teachers College Press.

National Committee on Test Standards. (1966). Standards for educational and psychological tests and manuals (pp. 12-24). Washington, DC: American Psychological Association. (Reprinted in Payne, D. A., & McMorris, R. F. (1967). Educational and psychological measurement: Contributions to theory and practice (pp. 76-80). Waltham, MA: Blaisdell.)

Parker, C. A. (1977). On modeling reality. Journal of College Student Personnel, 18, 419-425.

Pascarella, E. T. (1989). The development of critical thinking: Does college make a difference? Journal of College Student Development, 30, 19-26.

Pascarella, E. T. (1991). The impact of college on students: The nature of the evidence. Review of Higher Education, 14, 453-466.

Pascarella, E. T., & Terenzini, P. T. (1991). How college affects students: Findings and insights from twenty years of research. San Francisco: Jossey-Bass.

Patton, M. J. (1991). Qualitative research on college students: Philosophical and methodological comparisons with the quantitative approach. Journal of College Student Development, 32, 389-396.

Perry, W. G., Jr. (1968). Patterns of development in though and values of students in a liberal arts college: A validation of a scheme. Final report (DHEW Office of Education Project No. 5-0825). Washington, DC: U.S. Department of Health, Education, and Welfare. (ERIC Document Reproduction Service No. ED 024 315)

Perry, W. G., Jr. (1970). Forms of intellectual and ethical development in the college years: A scheme. New York: Holt, Rinehart and Winston.

Perry, W. G., Jr. (1978). Comments on chapter 2. In C. A. Parker (Ed.), Encouraging development in college students (pp. 60-63). Minneapolis, MN: University of Minnesota.

Perry, W. G., Jr. (1981). Cognitive and ethical growth: The making of meaning. In A. W. Chickering (Ed.), The modern American college (pp. 76-116). San Francisco: Jossey-Bass.

Rest, J., Cooper, D., Coder, R., Masanz, J., & Anderson, D. (1974). Judging the important issues in moral dilemmas--an objective measure of development. Developmental Psychology, 10, 491-501.

Rodgers, R. F. (1980). Theories underlying student development. In D. G. Creamer (Ed.), Student development in higher education: Theories, practices, and future directions (ACPA Media Publication No. 27, pp. 10-95). Cincinnati, OH: ACPA Media.

Rodgers, R. F. (1989). Student development. In U. Delworth & G. R. Hanson (Eds.), Student services: A handbook for the profession (2nd ed., pp. 117-164). San Francisco: Jossey-Bass.

Rodgers, R. F. (1990). Recent theories and research underlying student development. In D. G. Creamer (Ed.), College student development: Theory and practice for the 1990s (ACPA Media Publication No. 49, pp. 27-79). Alexandria, VA: American College Personnel Association.

Saidla, D. D. (1990). Roommates' cognitive development, interpersonal understanding, and relationship rapport. Journal of College Student Development, 31, 300-306.

Schmidt, J. A. (1985). Older and wiser? A longitudinal study of the impact of college on intellectual development. Journal of College Student Personnel, 26, 388-394.

Schmidt, J. A., & Davison, M. L. (1983). Helping students think. Personnel and Guidance Journal, 61, 563-569.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Spector, P. E. (1981). Research Designs. Sage University Paper series on Quantitative Applications in the Social Sciences (No. 07-001). Beverly Hills, CA: Sage.

Stage, F. K. (1989). College outcomes and student development: Filling the gaps. Review of Higher Education, 12, 293-304.

Stage, F. K. (1991). Common elements of theory: A framework for college student development. Journal of College Student Development, 32, 56-61.

Stage, F. K., & Russell, R. V. (1992). Using method triangulation in college student research. Journal of College Student Development, 33, 485-491.

Stephenson, B. W., & Hunt, C. (1977). Intellectual and ethical development: A dualistic curriculum intervention for college students. Counseling Psychologist, 6(4), 39-42.

Stonewater, B. B., Stonewater, J. K., & Hadley, T. D. (1986). Intellectual development using the Perry Scheme: An exploratory comparison of two assessment instruments. Journal of College Student Personnel, 27, 542-547.

Strange, C. (1983). Human development theory and administrative practice in student affairs: Ships passing in the daylight? NASPA Journal, 21(1), 2-10.

Taylor, M. (1983). New concepts in instrumentation development to measure the Perry Scheme. Paper presented at the meeting of the American College and Personnel Association, Houston, TX. (ERIC Document Reproduction Service No. ED 235 438)

Terenzini, P. T., & Pascarella, E. T. (1991). Twenty years of research of college students: Lessons for future research. Research in higher education, 32, 83-92.

Terenzini, P. T., Pascarella, E. T., & Blimling, G. S. (1996). Students' out-of-class experiences and their influence on learning and cognitive development: A literature review. Journal of College Student Development, 37, 149-162.

Walsh, W. B., & Betz, N. E. (1990). Tests and assessment (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Weiss, D. J., & Yoes, M. E. (1990). Item response theory. In R. K. Hambleton & J. N. Zaal (Eds.), Advances in educational and psychological testing: Theory and applications (pp.69-95). Boston: Kluwer Academic.

White, D. B., & Hood, A. B. (1989). An assessment of the validity of Chickering's theory of student development. Journal of College Student Development, 30, 354-361.

Widick, C. (1977). The Perry Scheme: A foundation for developmental practice. Counseling Psychologist, 6(4), 35-38.

Widick, C., & Simpson, D. (1978). Developmental concepts in college instruction. In C. A. Parker (Ed.), Encouraging development in college students (pp. 27-59). Minneapolis, MN: University of Minnesota.

Widick, C., Knefelkamp, L., and Parker, C. A. (1980). Student development. In U. Delworth & G. R. Hanson (Eds.), Student services: A handbook for the profession (pp. 75-116). San Francisco: Jossey-Bass.

Widick, C., Knefelkamp, L. L., & Parker, C. A. (1975). The counselor as a developmental instructor. Counselor Education and Supervision, 14, 286-296.

Wiersma, W., & Jurs, S. G. (1985). Educational measurement and testing. Boston: Allyn and Bacon.

Wilson, B. A. (1995). Intellectual development of technical college instructors. Journal of Vocational Education Research, 20(3), 29-50.

Wilson, B. A. (1995). The relationships between the intellectual development of technical college instructors and age, education, teaching experience, and supervisory experience. Delta Pi Epsilon Journal, 37, 95-106.

Wilson, B. A. (1996). A descriptive study: The intellectual development of business administration students. Delta Pi Epsilon Journal, 38, 209-221.

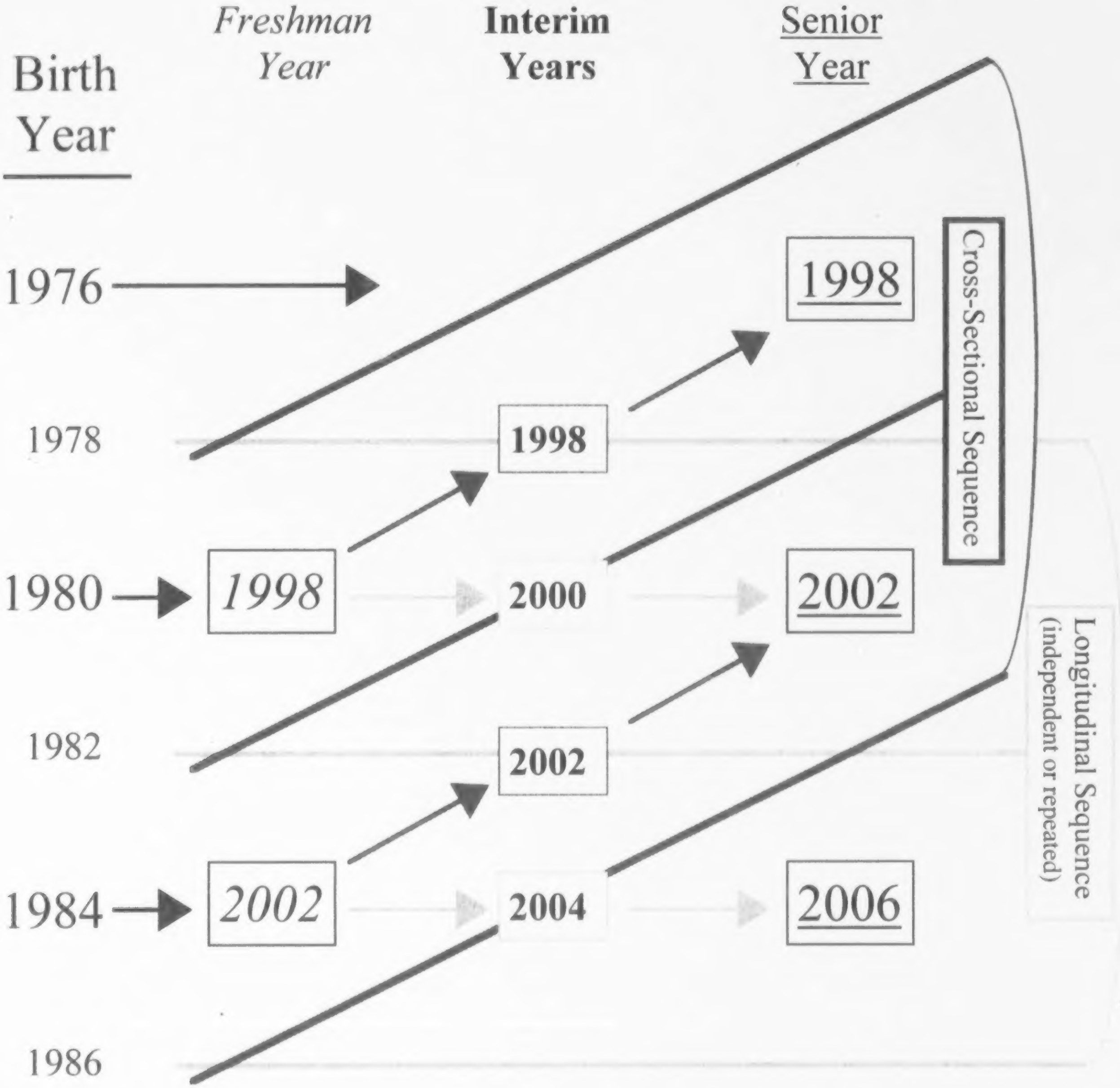
Winston, R. B., & Miller, T. K. (1994). A model for assessing developmental outcomes related to student affairs programs and services. NASPA Journal, 32(1), 2-15.

Winston, R. B., Jr., & Moore, W. S. (1991). Standards and outcomes assessment: Strategies and tools. In W. A. Bryan, R. B. Winston, Jr., & T. K. Miller (Eds.), New directions for student services: No. 53. Using professional standards in student affairs (pp.63-82). San Francisco: Jossey-Bass.

Wise, S. L. (1986). The use of ordering theory in the measurement of student development. Journal of College Student Personnel, 27, 442-447.

Figure 1

Sequential Designs



Birth Year
Freshman Year
Interim Years
Senior Year
1976
1978
1980
1982
1984
1986
1998
2002
1998
2000
2002
2004
1998
2002
2006
Cross-Sectional Sequence
Longitudinal Sequence (independent or repeated)

The dates (in boxes) in the interior of the figure indicate the year that the given age-rank group should be measured for each design. For example, for the Longitudinal Sequence repeated design: the 1980 age cohort should be measured in 1998 as freshmen with the same subjects measured again in 2002 as seniors; the 1984 cohort should be measured in 2002 as freshmen and 2006 as seniors.